UNIVERSITÉ MONTPELLIER II

ACADÉMIE DE MONTPELLIER

SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ MONTPELLIER II

Discipline : Virologie
Formation Doctorale : Biologie Santé
Ecole Doctorale : Science Chimique et Biologique de la Santé

DÉTECTION ET CARACTÉRISATION MOLÉCULAIRES RAPIDES DU VIRUS DE LA PESTE PORCINE AFRICAINE ET UTILISATION DES RECONSTRUCTIONS PHYLOGÉNÉTIQUES POUR RECONSTITUER SON HISTOIRE ÉVOLUTIVE

par

VINCENT MICHAUD

DIRIGÉE PAR MR. EMMANUEL ALBINA

Soutenue le 29 novembre 2012 devant le jury composé de :

MR. EMMANUEL ALBINA	DIRECTEUR DE RECHERCHE, CIRAD	Directeur de thèse
MR. CARLOS MARTINS	PROFESSEUR	Rapporteur
MR. MICHEL PETERSCHMITT	DIRECTEUR DE RECHERCHE, CIRAD	Rapporteur
MR. SERGE MORAND	DIRECTEUR DE RECHERCHE, CNRS	Examinateur
MR. GAËL THÉBAUD	CHARGÉ DE RECHERCHE, INRA	Examinateur
MME. MARYLÈNE TIGNON	DOCTEUR, CODA-SERVA	Examinateur

DÉTECTION ET CARACTÉRISATION MOLÉCULAIRES RAPIDES DU VIRUS DE LA PESTE PORCINE AFRICAINE ET UTILISATION DES RECONSTRUCTIONS PHYLOGÉNÉTIQUES POUR RECONSTITUER SON HISTOIRE ÉVOLUTIVE

RÉSUMÉ

La Peste porcine africaine (PPA) est une maladie contagieuse spécifique du porc domestique due au seul arbovirus à ADN identifié à ce jour. Décrite pour la première fois en 1921 au Kenya, la maladie a ensuite diffusé dans de nombreuses régions du monde. Malgré l'isolement de nombreuses souches virales au cours du temps, peu d'études phylogénétiques ont été menées jusqu'ici pour comprendre les relations unissant ces isolats entre eux. Or, la caractérisation est essentielle à la traçabilité des souches et donc à la compréhension de l'épidémiologie de la maladie. De plus, les conditions climatiques et environnementales des principaux pays atteints rendent difficiles l'accès, le transport et la conservation de nouvelles souches. Dans cette thèse, un protocole de prélèvement et de conservation du sang a été développé, pour la détection et la caractérisation rapides des souches. Une étude phylogénétique approfondie a été réalisée en utilisant des données de séquences publiques et inédites de virus isolés depuis 1950. Les analyses ont porté sur les gènes B646L, CP204L et E183L. Les analyses phylogénétiques ont utilisé les méthodes de maximum de vraisemblance et d'inférence bayésienne, qui ont permis de proposer une nouvelle nomenclature virale en 35 clusters différents. De plus, une datation des origines du virus a été menée, après avoir éliminé les biais d'analyse dus à une pression de sélection positive et/ou aux recombinaisons. L'horloge moléculaire a permis de déterminer que l'ancêtre commun le plus proche des souches contemporaines (TMRCA) se situait au début du 18ème siècle.

MOTS CLEFS : Peste porcine africaine – phylogénie – datation moléculaire – B646L – E183L – CP204L

RAPID MOLECULAR DETECTION AND CHARACTERIZATION OF AFRICAN SWINE FEVER VIRUS AND USE OF PHYLOGENETIC RECONSTRUCTIONS FOR EVOLUTIONARY HISTORY INFERENCE

ABSTRACT

African swine fever (ASF) is a highly lethal disease of domestic pigs caused by the only known DNA arbovirus. It was first described in Kenya in 1921 and since then a substantial number of isolates have been collected worldwide. However, only few phylogenetic studies have been carried out to better understand the relationships between isolates, which is essential for virus traceability and epidemiological understanding of the disease. Access, transport and virus conservation are also complicated by climatic and environmental conditions in affected developing countries. In this thesis, a simple method of blood sampling was developed allowing rapid virus detection and characterization. Comprehensive phylogenetic reconstructions were made using publicly and newly generated sequences of hundreds ASFV isolates of the last 60 years. Analyses focused on B646L, CP204L and E183L genes. Phylogenetic analyses were achieved using maximum likelihood and Bayesian coalescence methods and a new lineage based nomenclature is proposed to designate 35 different clusters. In addition, dating of ASFV origin was carried out from the molecular data sets. To avoid biased diversity, positive selection or recombination events were neutralized. The molecular clock analyses revealed that ASFV strains currently circulating have evolved over 300 years, with a time to the most recent common ancestor (TMRCA) going back to the early 18th century.

KEYWORDS: African swine fever – phylogeny – molecular clocking – B646L – E183L – CP204L

Laboratoire de Virologie, UMR CIRAD/INRA 1309 « Contrôle des maladies exotiques et émergentes », Campus International de Baillarguet, 34398 Montpellier Cedex 05 – France

REMERCIEMENTS

En premier lieu je tiens à remercier Carlos Martins et Michel Peterschmitt, qui ont accepté d'être les rapporteurs de cette thèse, ainsi que Serge Morand et Gaël Thébaud pour leur participation au jury. Je souhaite également remercier Stephan Zientara, Bernard Lebleu, Max Bergoin, Denis Fargette, Marie-Line Caruana, Stéphane Bertagnoli, Renaud Lancelot et Catherine Cêtre-Sossah d'avoir assisté aux comités de thèse qui ont jalonnés ces trois années, leur aide m'a été précieuse.

Cette thèse à une histoire. Nombreux sont ceux et celles qui ont participé à l'écrire. Parmi ces personnes, je voudrais en remercier ici particulièrement trois, par qui l'idée de cette thèse a été plantée, qui lui ont permis de germer, et puis qui l'ont accompagnée dans sa réalisation.

Tout d'abord, je tiens à remercier très chaleureusement Marie-Edith Lafon, qui m'a acceuilli lors de ma formation post-BTS durant une année, au laboratoire de Virologie du CHU Pellegrin de Bordeaux, entre 1995 et 1996. C'était il y a une petite éternité (j'ai appris depuis qu'il existe toutes sortes d'éternités), mais c'est avec elle que j'ai fait mes premiers pas dans le monde des virus. Et l'on n'oublie pas qui vous a transmis une passion.

Puis, j'exprime ma reconnaissance à Ultan Power, à qui je dois en premier lieu d'avoir appris mon métier de technicien de recherche en virologie. Durant les quatre années de notre collaboration, il n'eut de cesse de vouloir m'emmener « au-delà de la technicité ». J'ai essayé depuis de ne pas me soustraire à cette injonction aux accents irlandais.

Enfin, j'offre ma gratitude à Emmanuel Albina qui, après m'avoir recruté au CIRAD en 2002, m'a accompagné et soutenu depuis dix ans dans ce projet. D'abord pour l'obtention du diplôme de l'Ecole Pratique des Hautes Etudes en 2007, puis en acceptant d'être le Directeur de Recherche de ma thèse. J'espère répondre à la confiance qu'il m'a accordée.

Il va sans dire que je remercie aussi grandement ici tous ceux et toutes celles, collègues et ami(e)s, qui m'ont soutenu depuis ces années, et supporté depuis ces derniers mois. Une mention spéciale est attribuée à l'UMR15 du CIRAD à Montpellier, particulièrement à l'équipe du laboratoire de virologie. Avec un prix du jury pour Valérie, Armelle, Cécile, Laurence et Olivier. Merci aussi à mon amie Béatrice, ta présence m'est précieuse.

Puis, comment ne pas remercier la vie, qui a placé sur ma route ces personnes, et tant d'autres, qui font que j'écris ces lignes aujourd'hui. Aussi, à mes parents qui ont rendu cela possible, merci. A ma famille, Emmanuelle, Lola, Romane, Raphaël, merci.

Table des matières

INTRODUCTION

Les maladies émergentes à fort impact sanitaire sont pour la plupart d'origines virales et affectent aussi bien les humains que les animaux (voire les deux), ainsi que les plantes. Ces émergences résultent majoritairement de modifications d'équilibres consécutifs à l'activité humaine : anthropisation de nouveaux écosystèmes favorisant le contact avec de nouveaux réservoirs d'agents pathogènes, modifications environnementales, comme le changement climatique, l'érection de nouveaux barrages ou la déforestation. Une émergence dépend également de la plasticité du génome microbien permettant le franchissement de la barrière d'espèce et une meilleure adaptation à l'hôte. La plasticité des génomes viraux peut être analysée en utilisant des méthodes mathématiques d'analyse de séquences décrivant les modalités sous-jacentes de génération de diversité. De la même façon, les relations entre isolats dans l'espace et le temps peuvent être inférées.

L'objet de ce travail porte sur la détection, la caractérisation et l'analyse phylogénétique approfondies d'un virus d'importance en santé animale, le virus responsable de la Peste porcine africaine (PPA), grand virus à ADN génomique double brin et à transmission directe ou vectorielle (tique). Depuis son émergence rapportée au siècle dernier, la PPA est en expansion accélérée depuis une vingtaine d'années, probablement en raison des changements globaux résultant des activités humaines et ayant favorisé le trafic de matières virulentes. La phylogénèse du virus de la PPA reste encore mal connue aujourd'hui, mais quelques études phylogénétiques ont été menées depuis une décennie.

Dans ce manuscrit, avant de présenter nos résultats, nous rappellerons en premier lieu les méthodes qui ont permis de caractériser et de classer la diversité du vivant au cours du temps, partant des conceptions philosophiques à la modélisation mathématique opérationnelle des processus évolutifs. Les modèles mathématiques ont été introduits pour décrire et représenter notamment sous forme d'arbres, les relations entre organismes. La détermination et l'analyse des séquences offrent la possibilité d'une classification individuelle des organismes, c'est-à-dire la construction d'un arbre permettant de positionner chacune de ses feuilles et de les connecter entre elles au moyen de branches évolutives. Cependant, dans les arbres représentant la diversité, les virus tiennent une place à part. En effet, les relations entre les différents clusters de l'arbre du vivant sont basées sur l'analyse de gènes ribosomaux qui sont absents chez les virus. De plus, l'on ne peut être certain de l'origine des gènes viraux, à savoir s'ils sont le produit d'une histoire évolutive issue d'une spéciation, de recombinaisons entre deux virus ou s'ils ont été acquis de leur(s) hôte(s) au fil du temps. Les virus possèdent donc un rythme évolutif propre. Les méthodes décrivant l'évolution des séquences nucléotidiques sont toutefois applicables à l'étude de l'évolution des virus. Après avoir exposé le grand principe de ces méthodes, nous décrirons

l'état actuel des connaissances concernant l'évolution des virus, que leur génome soit constitué d'ADN ou d'ARN, simple ou double brins.

Dans un second temps, après avoir exposé l'état de l'art sur les relations phylogénétiques entre isolats de virus PPA, nous présenterons la nature et les objectifs de cette thèse. Enfin, nous présenterons nos travaux en trois parties, portant sur la détection et la caractérisation de virus PPA en condition africaines, l'intérêt des méthodes de reconstruction phylogénétique pour identifier les sources de contamination et enfin, nos analyses approfondies sur l'évolution du génome viral. Une discussion générale portant sur l'ensemble de nos résultats sera proposée en fin de manuscrit.

1- L'évolution du vivant

1-1- Historique de la classification des espèces

1-1-1- La classification des espèces de l'antiquité au siècle des lumières

L'Homme s'est semble-t-il toujours interrogé sur les liens qui existent entre l'observation de ce qui l'entoure et la connaissance « vraie », que ce soit celle des espèces animales, végétales, ou minérales. Aristote (384 – 323 av. J.C.) a vraisemblablement été l'un des tous premiers, si ce n'est le premier philosophe, à avoir compris que le principe même de toute connaissance reposait sur la rigueur avec laquelle on l'appréhendait et on la retranscrivait. Dans sa « *Métaphysique »*, Le philosophe jette les bases de la méthode présidant à toute avancée véritablement scientifique : la connaissance des causes, ou principes, et, par la même, de la nécessité des choses et des êtres, c'est-à-dire des effets induits par ces causes. Ces principes se devaient d'être débarrassés du mysticisme dont les entouraient ses prédécesseurs, tel Platon, pour être basés seulement et uniquement sur des faits observables. Avec sa « *Parva naturalia »* (Petits traités d'histoire naturelle), et surtout ses « *Parties des animaux »* et « *Génération des animaux »* (343 av. JC), il est considéré comme le premier naturaliste. Au travers de ces œuvres, il tente une classification descriptive compréhensible des animaux en espèces et en genres. Il refuse la théorie de l'évolution que Démocrite (460 – 370 av. JC) avait formulée, basée sur le hasard de la rencontre et de l'assemblage des atomes, en prônant l'existence éternelle des choses et des êtres : c'est la théorie *fixiste*. Cependant, la classification systématique trouvait là ses fondements, selon une approche encyclopédique, basée sur la logique pure. Ces fondements perdurèrent jusqu'à la fin du moyen âge, période durant laquelle classer, et donc reconnaitre, les espèces, notamment végétales, pouvait s'avérer crucial en termes de survie de l'Homme (pensons par exemple, aux plantes vénéneuses devant être différenciées des plantes comestibles).

Deux naturalistes contemporains contribuèrent ensuite à la naissance du système moderne de la classification des espèces ainsi qu'à la mise en exergue des liens qui les unissent : le français Georges – Louis Leclerc de Buffon (1707 – 1788) et le suédois Carl Von Linné (1707 – 1778). C'était le début du siècle des lumières. Le premier s'intéressa tout particulièrement aux plantes, en tant qu'intendant du jardin du roi, et basa son système de classification sur l'anatomie comparée des espèces vivantes. Il parvint à imaginer une théorie du *transformisme* dans laquelle la morphologie et le devenir des espèces tenaient de leur environnement. Cette théorie prenait en compte la comparaison entre les espèces, jetant ainsi les prémisses de la théorie de l'évolution qui serait développée au cours du XIXème siècle, en imaginant un ancêtre commun possible aux espèces les plus semblables. Carl Von Linné, quant à lui, posa les termes d'une nomenclature des espèces vivantes, appelées taxa, dans son œuvre majeure, le « *Systema naturae »* (1735), dont il fera plusieurs rééditions. Dans la dernière de ces rééditions, en 1758, il établit une nomenclature binomiale définitive, en attribuant de façon systématique des noms en latin (la langue universelle des savants de cette époque), à toutes les espèces animales. Il répertoriera également plus de 8000 espèces de plantes dans son livre « *Species plantarum »*, en 1753, selon la même nomenclature latine. L'organisation de sa classification tenait compte de sept niveaux hiérarchiques de classement des êtres vivants : il s'agissait du règne, de l'embranchement, de la classe, de l'ordre, de la famille, du genre et de l'espèce. Tout comme Aristote, Linné était *fixiste*, et, si le premier pensait un dieu en tant qu'entité primordiale, unité originelle (monade), cause ultime de tout, le second pensait Dieu comme origine de l'organisation de toute chose, et donc à l'immuabilité de son œuvre. La pensée créationniste linnéenne liée à Dieu ne permit pas à sa classification de perdurer. Cependant, sa nomenclature rigoureuse passa, elle, à la postérité. Elle est, de fait, toujours employée aujourd'hui par les systématiciens.

Fixistes et *transformistes* s'opposèrent au cours des XVIIIème et XIXème siècles. Parmi les plus célèbres controverses, on trouve celle ayant opposée les naturalistes français Georges Cuvier (1769 – 1832) et Jean-Baptiste de Lamarck (1744 – 1829). Cuvier consacra l'essentiel de son travail à l'anatomie comparée des espèces, en étant persuadé de leur fixité, l'origine divine du vivant étant pour lui un préalable à toute chose. Cependant, également géologue et paléontologue, il constata la possibilité de l'extinction des espèces au cours des temps géologiques, qu'il décrivit dans son ouvrage « *Les révolutions de la surface du globe »* (1825). Lamarck, de son côté, échafauda la première théorie de l'évolution. Il décrivit un processus naturel au cours duquel les espèces vivantes, mues par la pression de leur environnement (ou contingences), étaient contraintes de se transformer, et pouvaient léguer ces transformations à leur descendance. Les contingences environnementales n'évoluant jamais vers la simplification, le processus de transformation devait lui aussi évoluer vers toujours davantage de complexité. Lamarck développera cette théorie *transformiste*, en se basant sur sa connaissance des mollusques, au cours de « l'*Histoire naturelle des animaux sans vertèbres »* (1815 – 1822). Cette théorie postulait que la nature avait produit tous les êtres vivants, partant du plus simple pour parvenir au plus complexe,

et que les contingences environnementales auxquelles ces êtres avaient été confrontés expliquaient les différences constatées. Cependant, comme tous les scientifiques de cette époque, avant les travaux de Pasteur, Lamarck accepte l'idée de génération spontanée à partir de la matière. En définitive, il ne réussit jamais à imposer sa théorie, mise en défaut, notamment, par l'absence, soulignée par Cuvier, de formes intermédiaires entre les différents embranchements des règnes du vivant, les fameux chainons manquants.

Avant les travaux de Pasteur, la plupart des scientifiques acceptaient l'idée de génération spontanée, c'est-à-dire l'apparition des êtres vivants n'ayant pas d'ascendants. Cette théorie, prend sa source dans l'antiquité, Aristote en faisant déjà la synthèse en compilant les écrits de ces prédécesseurs. L'invention du microscope, au XVIème siècle, par Antoni Van Leeuwenhoek (1632 – 1723) permit d'accéder à l'observation du monde infra-visible, c'est-à-dire unicellulaire et microbien. La découverte et la caractérisation qu'il fit de bactéries responsables de certaines maladies humaines permit à Pasteur de réfuter la théorie dite de « *l'hétérogénie* », en prouvant qu'à l'origine de tous les cas dits de génération spontanée, se trouvait des germes ou des œufs ayant permis l'éclosion de nouveaux êtres vivants. Désormais, la classification du vivant prendrait en compte la théorie microbienne ou cellulaire.

1-1-2- La théorie de l'évolution, ou la sélection naturelle

C'est en 1859 qu'un bouleversement intellectuel se produisit dans le monde scientifique de l'époque, avec la publication par Charles Darwin (1809 – 1882) de « *L'origine des espèces* ». Cette publication fut à l'origine d'une révolution comparable aux théories galiléennes ou coperniciennes. Darwin ne fut d'ailleurs pas le seul scientifique de son temps à postuler l'évolution des espèces par la sélection naturelle. Alfred R. Wallace (1823 – 1913), son contemporain, la décrivit lui aussi dans son ouvrage « *De la tendance des variétés à s'écarter indéfiniment du type primitif* », également paru en 1859. C'est cependant « *L'origine des espèces* » de Darwin qui fera date dans l'histoire de l'évolution. Il est à noter que le terme histoire ne doit pas être entendu dans son acception d'évènements chronologiques, mais au sens de la quête d'informations pouvant expliquer, étayer, une observation.

La théorie décrite dans *L'origine des espèces* se base sur le développement de deux grands thèmes, issus des constats de l'auteur, et qui visent à expliquer à la fois, la répartition des êtres vivant sur la terre, êtres vivants appartenant tant au monde animal que végétal, que leurs apparitions et présences successives en ces mêmes lieux. Le premier constat qu'effectue Darwin est que les descendants de parents uniques présentent des caractéristiques différentes qu'ils peuvent cependant transmettre à leur propre descendance. Le second constat est que les descendances sont impliquées dans une lutte sans merci pour la survie. En effet, la descendance d'une espèce devrait s'accroître selon

une progression géométrique, rendant impossible leur coexistence, ne serait-ce que pour des raisons d'accès aux ressources autorisant la survie de chacun. A chaque génération, une partie plus ou moins importante est donc éliminée, entre autre pour que survive le reste de la communauté.

Darwin postule donc le fait que chaque espèce, et chaque individu qui la compose, est en lutte pour s'octroyer les moyens de sa subsistance, de son existence. Les différences phénotypiques bénéfiques pour un individu donné, dans un environnement donné, seront donc avantagées et préservées au fil de leurs descendances, assurant ainsi leur pérennité. C'est cette sélection naturelle qui, répétée au fil des générations successives, assure la diversité et la pérennité du vivant que l'on peut constater par la seule observation.

Cependant, si la sélection naturelle parvenait à la génération d'un être parfait, c'est-à-dire parfaitement adapté à son environnement, l'évolution n'aurait alors plus aucune raison de se poursuivre. Darwin rejette d'emblée cette hypothèse, en arguant du fait que la perfection n'est pas le but de l'évolution, et que seule une variation même minime, mais bénéfique, permet sa pérennisation puisqu'elle offre un avantage substantiel dans la lutte pour la vie. Il nie donc la possibilité de « saut évolutif » pour lui préférer la notion de progressivité, de continuité du changement. Cette notion avait d'ailleurs déjà été développée par Aristote, reprise par Gottfried W. Leibniz (1646 – 1716) dans ses « *Nouveaux essais sur l'entendement humain* » (1704), puis par Linné dans son « *Philosophia botanica »* en 1751. Cette notion, qui jusqu'alors ne trouvait sa source que sur un plan philosophique, a tenté depuis d'être étayée biologiquement. La discipline scientifique la plus probante ayant pu venir au secours de cette théorie de l'absence de saut évolutif, c'est-à-dire du principe de continuité, a été sans conteste l'embryologie. L'embryologie a, en effet, montré que l'embryon de tous les animaux était formé, lors de son stade initial, de trois feuillets à partir desquels tous les organes se développent (Baer 1827). Par suite, l'embryologie comparée a montré que de grandes analogies existaient entre les premiers stades du développement embryonnaire d'espèces mêmes très éloignées dans l'arbre du vivant. Ainsi, la théorie de la *récapitulation* énoncée par Ernst Haeckel en 1866 a établi durant des décennies, que le développement de l'embryon (ontogénèse) gardait les traces du passé évolutif de l'espèce à laquelle il appartient (phylogénèse). C'est à partir des années 1970 que cette théorie de la continuité évolutive a été battue en brèche, notamment par l'apparition de la théorie des « *équilibres ponctués* » énoncée par Stephen Jay Gould et Niles Eldredge, en 1972. Cette théorie nouvelle postule que l'évolution des êtres vivants, si elle est, pour l'essentiel, constituée de périodes durant lesquelles l'évolution suit un rythme lent (c'est-à-dire comparable au principe de continuité), par le jeu des mutations et de la sélection naturelle, est cependant sujette à des accélérations évolutives ponctuelles et brèves, caractérisées par des extinctions ou des spéciations massives. Cette théorie ne remet donc pas complètement en cause la théorie de l'évolution darwinienne, mais vient davantage la compléter, en évacuant le problème du chainon manquant, être vivant représentant une mosaïque entre

deux espèces proches mais disjointes, et qui devrait systématiquement exister si le principe de continuité évolutive était absolument vrai.

Nous l'avons vu, philosophie et science de l'évolution furent longtemps intimement liées, en l'absence de méthodes permettant une observation des différences entre individus qui ne soit pas seulement phénotypiques, anatomiques ou comportementales.

1-2- La transmission génétique de l'hérédité : de Mendel à la découverte de l'ADN

Si la théorie de la sélection naturelle décrite par Darwin s'appuie sur l'observation de la diversité du vivant, elle recèle une lacune, car elle n'explique pas les forces évolutives biologiques sous-jacentes à l'origine de cette diversité. Nous l'avons vu, la théorie de la sélection naturelle prend en compte les caractères transmis entre les générations, mais sans pouvoir expliquer les mécanismes ni le support de cette hérédité. De même, aucune modélisation mathématique ou statistique ne vient la supporter. C'est donc une théorie implicite plus qu'explicite.

Le premier à avoir tenté de comprendre la façon dont les caractères parentaux étaient transmis aux générations postérieures fut sans conteste Gregor Mendel, moine de son état (1822 – 1884). Il mit à jour les lois de la génétique, qui porteront plus tard le nom de *génétique mendélienne*, et qui fixent la façon dont les gènes sont transmis de générations en générations. Botaniste, il s'était donné pour objectif de comprendre le phénomène d'hybridation chez les végétaux. Pour ce faire, il croisa, pendant plus de dix années, des plans de *Pisum sativum*, ou Pois cultivé (une espèce de légumineuse de la famille des Fabacées), et observa scrupuleusement comment les phénotypes parentaux étaient transmis et répartis chez leurs descendances.

A partir de ces observations, il émit trois lois fondamentales qui ont toujours force de loi aujourd'hui (Mendel 1866). La première de ces trois lois stipule qu'à partir de parents de souche pure F0 (que l'on nommera homozygotes ultérieurement), il y a uniformité de la descendance hybride F1. La seconde loi postule que les gamètes sont le siège des facteurs héréditaires, ne contenant qu'un seul des caractères phénotypiques de l'individu ; la ségrégation des caractères phénotypiques apparaît donc à la génération F2. Enfin, la troisième loi énonce que la ségrégation des caractères héréditaires est indépendante en génération F2, c'est-à-dire que les allèles sont disjoints. Les travaux de Mendel ne furent pas reconnus en leur temps, notamment par Darwin qui ne les prit jamais en considération pour étayer sa propre théorie, jusqu'à ce que les lois gouvernant l'hérédité soient redécouvertes simultanément par trois botanistes, au début du XXème siècle : H. de Vries (néerlandais), C.E. Correns (allemand) et E. von Tschermark (autrichien). Ces lois restaient cependant basées sur l'observation, sans qu'elles n'aient été jusqu'alors modélisées mathématiquement. Elles servirent néanmoins de base à des concepts statistiques tels que

le maximum de vraisemblance, introduit par Ronald Aylmer Fisher (1890 – 1962) dès 1922. Fisher fut d'ailleurs, avec J.B.S. Haldane à l'origine de la *génétique des populations*, méthode mathématique qui vise à comprendre comment les différents allèles d'un même gène se répartissent au sein d'une population, et pouvant conduire au phénomène de spéciation (Fisher 1930 ; Haldane 1932).

Depuis l'invention du microscope au XVIème siècle, le perfectionnement du matériel d'observation avait permis de pénétrer l'intérieur de la cellule et d'observer les organites intracellulaires. C'est ainsi qu'en 1882 l'allemand W. Flemming (1843 – 1905) découvrit la migration polaire au cours de la mitose cellulaire d'un matériel nucléaire appelé jusqu'alors nucléine, et qu'il nomma chromosomes. De même, en 1902, W. Johannsen (1857 – 1927) comprit que les informations sur les caractéristiques phénotypiques transmises lors de la reproduction devaient être contenues à l'intérieur des cellules, les gamètes étant unicellulaires. Il nomma « gènes » ces éléments informatifs.

Jusqu'au début du XXème siècle, c'est donc à partir de végétaux, plus simples d'utilisation, que les lois de l'hérédité avaient été formulées. Il restait à vérifier leur réalité dans le règne animal. Pour ce faire, l'organisme modèle sélectionné fut la *Drosophila melanogaster*, ou mouche du vinaigre. Le choix d'un insecte tel que la drosophile fut dicté notamment par un temps de reproduction bref permettant l'observation de nombreuses générations. C'est Thomas Hunt Morgan (1866 – 1945), embryologiste et généticien, qui s'attacha à étudier les variations phénotypiques chez cette mouche. C'est ainsi qu'il fit le lien entre la transmission de caractères isolés, se rapportant aux éléments informatifs « gènes », et la recombinaison des chromosomes, lors de la division cellulaire. Il conclut ainsi que les chromosomes étaient en fait composés de nombreux gènes, eux-mêmes supports de l'information phénotypique transmise entre les générations successives.

Restait à comprendre comment l'information contenue dans les gènes était traduite en termes de phénotype chez un individu donné. Ce sont deux américains, G.W. Beadle et E.L. Tatum qui découvrirent que l'information contenue dans un gène était traduite en protéine (Beadle & Tatum 1941), sans pour autant décrypter la nature biochimique des gènes. La composition biochimique des gènes fut révélée trois ans plus tard, en 1944, par O.T. Avery qui découvrit que la nucléine, découverte chez les bactéries, était composée d'acide désoxyribonucléique, ou ADN (Avery OT 1944). La composition chimique de l'ADN est faite de seulement quatre éléments, des bases azotées : l'adénine (A), la guanine (G), qui sont des bases puriques, et la cytosine (C) et la thymine (T), des bases pyrimidiques (Figure 1). E. Chargaff, en 1950, détermine les proportions paritaires entre A et T d'un côté et G et C de l'autre, dans la molécule d'ADN : soit A/T = 1 et G/C = 1 (Chargaff *et al.* 1950). La structure spatiale définitive de l'ADN sera mise à jour par James D. Watson et Francis H.C. Crick en 1953 (Watson & Crick 1953) : les quatre nucléotides A, T, C et G s'enchainent sur deux brins antiparallèles appariés formant une double hélice dextrogyre. Les liaisons qui unissent ces deux brins s'établissent entre adénine et thymine par deux liaisons hydrogènes, et entre

guanine et cytosine par trois liaisons hydrogènes. La complémentarité entre A – T et C – G expliquant de fait les ratios qu'avait déterminés Chargaff.

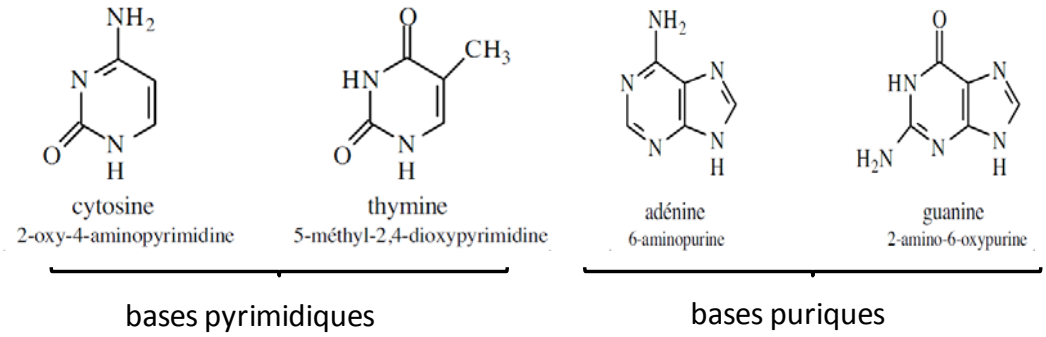


Figure 1 : Représentation de la formule chimique des quatre nucléotides composant la molécule d'ADN

Les quatre nucléotides qui s'enchainent peuvent donc être comparés à un alphabet composant des mots, les gènes, et c'est la lecture de ces mots qui permet à l'information contenue dans l'ADN d'être traduite en protéine. C'est Crick, en 1957, qui détermina que le passage du gène d'ADN à la protéine était réalisé grâce à une molécule intermédiaire : l'acide ribonucléique, ou ARN (Crick *et al.* 1957). Il est à noter que, dans l'ARN, la thymine est remplacée par un uracile (U). Ces trois molécules, ADN – ARN – protéine devaient devenir le dogme central de toute la biologie moléculaire qui s'est depuis développée. Les analyses chimiques des protéines avaient déterminé que leur composition était faite de 20 molécules différentes, les acides aminés, et une dernière découverte clé pour la compréhension de la diversité, fut la façon dont l'enchainement des bases dans l'ADN permettait celui des acides aminés dans la protéine. C'est ainsi que le code génétique fut mis au jour, en déterminant que les nucléotides étaient lus par les ribosomes par groupe de trois, ces triplets étant appelés des codons, et que chaque codon était traduit par un acide aminé (Figure 2).

Première lettre (côté 5')	Deuxième lettre U	Deuxième lettre C	Deuxième lettre A	Deuxième lettre G	Troisième lettre (côté 3')
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G
	codon d'initiation		codon de terminaison		

Figure 2 : Le code génétique universel. Le tableau est à triple entrée, chaque entrée représente la place du nucléotide dans le codon. Les trois codons « stop » sont indiqués en rouge, tandis que le codon initiateur, codant pour le premier acide aminé de toute protéine (AUG, codant pour la méthionine) est indiqué en vert.

Le code génétique est pour ainsi dire universel tant chez les eucaryotes (organismes dont les cellules possèdent un noyau « vrai ») que chez les procaryotes (organismes unicellulaires sans noyau). Les rares exceptions à cette universalité concernent les génomes particuliers qui se trouvent dans les mitochondries (censées être d'anciennes bactéries phagocytées puis devenues symbiotiques dans les cellules d'organismes eucaryotes) et dans les génomes d'un groupe particulier de bactéries du genre *Mycoplasma* (Osawa 1995).

Même si le code génétique est universel, tous les organismes ne l'utilisent pas de la même manière. En comparant les séquences géniques de nombreux organismes vivant dans des environnements différents, il a été établi une « préférence » d'utilisation entre des codons synonymes, basée sur la fréquence à laquelle ces organismes les utilisent. Il a été déterminé que cette préférence codon s'appliquait, telle une pression de sélection, sur tous les gènes d'un même organisme (Grantham *et al.* 1980), plus drastiquement encore lorsqu'il s'est agit de gènes fortement exprimés (Grantham *et al.* 1981). Cette utilisation préférentielle, appelée « biais d'usage de codons », pourrait jouer un rôle important en termes d'histoire évolutive des organismes (Ingvarsson 2008). De fait, la préférence codon d'un organisme reflète la composition en ARNt à l'intérieur de ses cellules. Les ARNt les plus nombreux étant plus souvent utilisés lors de la traduction, les gènes présentant ses codons possèderont donc un avantage en termes d'expression. En tant que parasites obligatoires utilisant la machinerie cellulaire pour se répliquer, un virus est aussi soumis au biais d'usage propre à l'hôte qu'il infecte. Les gènes viraux présentant les codons préférentiels seront donc favorisés, induisant l'émergence des virus les mieux adaptés à leur hôte (Cheng *et al.* 2012 ; Ma *et al.* 2011).

1-3- Le support de la diversité : la molécule d'ADN

1-3-1- Les signes de la diversité

1-3-1-1- Les substitutions, les insertions – délétions (indels)

Nous venons de le voir, toute expression phénotypique d'un organisme tire son origine de la composition nucléotidique de son génome et donc de l'expression de ses gènes, dont la séquence nucléotidique, c'est-à-dire la composition des codons, lui est propre.

Du fait de l'existence de quatre nucléotides, il y a 4^3, soit 64 combinaisons possibles de codons différents. Sur ces 64 combinaisons, trois ont pour fonction d'arrêter la traduction en protéine de l'ARN par les ribosomes ; ce sont les codons « stop » : UAA, UAG et UGA. Il reste donc 61 combinaisons ayant un sens biologique, c'est-à-dire codant réellement pour un acide aminé. Or, nous l'avons vu, il n'existe seulement que 20 acides aminés composant les protéines. Cela signifie que plusieurs codons codent pour un même acide aminé. C'est la raison pour laquelle le code génétique est dit « dégénéré », ou redondant. Sur ces 61

combinaisons, deux codons échappent à cette règle, ne codant chacun que pour un seul acide aminé. Il s'agit du codon AUG, ou codon d'initiation, codant pour la Méthionine, et du codon UGG, qui code pour le Tryptophane. Le codon AUG est dit d'initiation car c'est par lui que débute la traduction de toutes les protéines. En règle générale, cette Méthionine initiale est modifiée biochimiquement à la suite de la traduction, et même, la plupart du temps, est excisée de la protéine finalisée post-traductionnelle. Parmi les 59 combinaisons de codons restantes, on trouve donc de nombreux codons que l'on nomme synonymes, puisque codant pour un même acide aminé. Par exemple, la Sérine est codée par pas moins de sept codons différents.

Les acides aminés, de par leur nature (hydrophyle, hydrophobe, ou électriquement neutre), leur conformation dans l'espace, et les radicaux qu'ils portent, ont une importance fondamentale dans la structure et la fonction de la protéine finale néo-synthétisée. La conformation des protéines étant à la base des caractéristiques phénotypiques d'un organisme, la dégénérescence du code génétique a alors un impact non négligeable sur le concept de diversité. En effet, nous l'avons vu, l'origine de toute diversité prend sa source dans la composition nucléotidique de la molécule d'ADN formant les gènes. Toute modification dans la séquence d'un codon, si chaque codon n'encodait qu'un seul acide aminé aurait donc des conséquences immédiates en termes de protéines traduites. Ainsi, la dégénérescence du code génétique limite l'impact des substitutions nucléotidiques dans la séquence d'ADN, qui porte en elle la véritable diversité entre les organismes.

On note cependant que les substitutions n'ont pas les mêmes conséquences selon la position du nucléotide concerné dans le codon. En effet, on constate que les deux premières positions dans le codon ont une importance fondamentale. Toute modification du nucléotide en position 2 du codon implique systématiquement un changement de l'acide aminé ; c'est ce que l'on nomme une substitution non synonyme (Kimura 1983 ; Yang 1996), souvent caractéristique d'une évolution sous pression de sélection purificatrice, donc lente (Xia 1998). En revanche, la plupart des substitutions en position 3 sont synonymes (ou silencieuses) et au moins quelques-unes des mutations en positions 1 le sont également (Kimura 1977). On peut donc classer l'impact des substitutions dans le codon comme suit : $\mu_2 > \mu_1 > \mu_3$, où μ_n est le taux de substitution à chaque site dans le codon.

Outre les mutations synonymes et non synonymes, les mutations nucléotidiques peuvent être de deux types : les transitions, qui sont le remplacement d'une base purique par une autre base purique (A – G) ou une pyrimidine par une autre pyrimidine (C – T), et les transversions, où une base purique est remplacée par une base pyrimidique et inversement (A – C, T – G, A – T et C – G), soit un total de six substitutions possibles sur chaque site d'une séquence d'ADN (Figure 3). Bien que les possibilités de transversions soient trois fois plus nombreuses que les transitions, leur équiprobabilité n'est pas réalisée, puisqu'il a été observé que les transitions étaient nettement plus fréquentes au cours de la réplication de l'ADN (Fitch 1967 ; Gojobori *et al.* 1982).

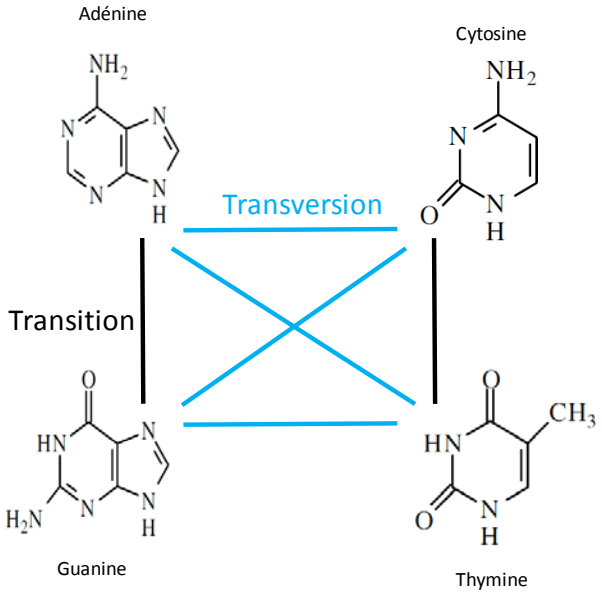


Figure 3 : Substitutions nucléotidiques pouvant se produire dans une séquence d'ADN. Quatre transversions entre bases puriques et pyrimidiques (A/G vs T/C) et deux transitions entre bases de même nature, pyrimidines (T vs C et A vs G), soit six possibilités mutationnelles.

Les substitutions se produisent le plus souvent spontanément lors de la réplication de l'ADN génomique. L'un des principaux mécanismes de leur apparition est dû à ce que l'on nomme les « *différences tautomériques* » (Wang *et al.* 2011). Les nucléotides, en effet, coexistent sous forme de tautomères, c'est-à-dire d'analogues structuraux (isomères) qui sont en équilibre dynamique dans la cellule. A titre d'exemple, la thymine se décline en deux tautomères, les formes *keto* et *enol* (Figure 4). La forme *keto* est la plus stable, avec une double liaison entre le carbone 4 de son cycle aromatique et un atome d'oxygène tandis que dans la forme *enol*, c'est un groupement OH qui est lié à ce même carbone, donc par une liaison simple. La forme *keto* est la plus facilement incorporée dans la molécule d'ADN néo-synthétisée, et se liera avec une adénine, tandis que si c'est la forme *enol* qui est incorporée, elle s'appariera avec une guanine, introduisant de la diversité dans les molécules d'ADN filles ultérieures. Cette propriété des ADN polymérases à ne pas pouvoir discerner aisément des analogues nucléosidiques a d'ailleurs été largement utilisée dans la lutte antivirale contre des virus comme le HIV ou le virus de l'hépatite C, en utilisant des nucléotides modifiés ayant un effet terminateur de polymérisation de l'ADN (Hamers *et al.* 2012 ; Pawlotsky 2012). Malgré les possibilités d'incorporation erronée de nucléotides lors de la synthèse du brin d'ADN fille, les polymérases sont des enzymes très fidèles à leur matrice, non seulement grâce à leur fidélité intrinsèque, mais aussi grâce aux enzymes de détection et de réparation des erreurs, présentes dans les cellules. Il a été estimé que ce système de sauvegarde de l'intégrité originelle d'une séquence d'ADN lors de sa réplication permettait un taux d'erreur situé entre 10^{-9} et 10^{-10} mutations par base répliquée (Echols & Goodman 1991).

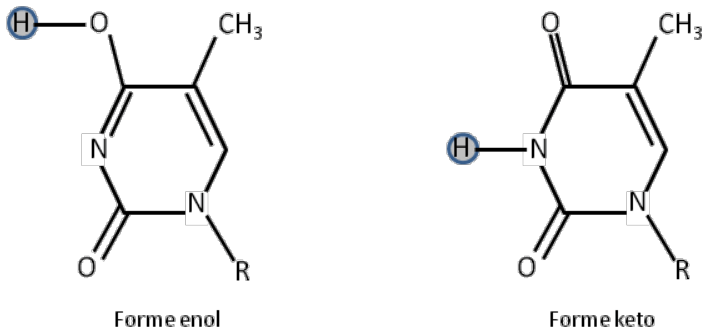


Figure 4 : Représentation des deux tautomères de la thymine. La forme *keto*, de part la double liaison entre son carbone 4 et un oxygène, est la plus stable. La forme *enol* lorsqu'elle est incorporée dans l'ADN, s'apparie avec une guanine plutôt qu'avec une adénine, introduisant de la diversité entre les séquences.

On peut noter, à titre informatif, que d'autres mécanismes peuvent induire des substitutions nucléotidiques, tels que des agents mutagènes, comme par exemple l'éthylméthane sulfonate, qui provoque des alkylations de la guanine qui s'apparie alors avec une thymine au lieu d'une cytosine, ou encore des rayonnements tels que les ultra-violets qui induisent des dimérisations nucléotidiques. Nous n'entrerons pas plus avant dans ces mécanismes, car ce n'est pas ici le but de notre propos.

L'observation des divergences entre des séquences d'ADN ne repose pas seulement sur la caractérisation des différences nucléotidiques (transitions et transversions) qui se sont accumulées entre elles au fil du temps. Les insertions et les délétions, également appelées indels (Zhang & Gerstein 2003), si elles sont moins fréquemment constatées au sein des séquences génomiques codantes que les substitutions nucléotidiques, représentent cependant une trace évidente d'évolution différentielle individuelle et d'accumulation importante de divergence (Tian *et al.* 2008). Les indels peuvent prendre plusieurs formes et avoir différentes conséquences pour la protéine codée par le gène concerné.

La structure tertiaire d'une protéine, c'est-à-dire la structure définitive qu'elle adopte en fonction des acides aminés qui la compose (structure primaire) et des modifications post-traductionnelles dont elle sera l'objet peut être affectée de façon dramatique par la présence d'indels. Il a toutefois été observé que la topologie des protéines pouvait être conservée quand bien même jusqu'à 70% de leur séquence pouvait être différente (Grishin 1997). Ceci souligne que la topologie d'une protéine est nettement plus conservée que sa séquence ou même que sa fonction (Zhang *et al.* 2012). Cependant, si, dans une séquence d'ADN génomique codante, les indels ne forment pas des multiples de trois nucléotides, c'est-à-dire d'unités de lecture de la séquence par le ribosome (Codoner *et al.* 2005), il se produit un changement du cadre de lecture du gène entrainant généralement la production d'une protéine radicalement différente de la protéine originelle. Outre la perte de conformation et de fonction qui s'ensuivent, si le devenir de la molécule est d'être une protéine structurelle, un tel changement peut s'avérer létal pour l'organisme. Si tel est le

cas, ces modifications, si elles font partie des processus évolutifs, ne participeront pas à l'évolution puisqu'elles ne peuvent être transmises à une quelconque descendance, l'organisme concerné étant dans l'impossibilité d'en avoir une. Les indels non létaux sont cependant considérés comme des marqueurs génétiques important au sein de populations naturelles, car leur présence est généralement transmise à la descendance (Vali *et al.* 2008).

Les indels peuvent également prendre la forme de séquences répétées de plusieurs codons, et donc de plusieurs acides aminés dans la séquence protéique traduite. Les mécanismes à l'origine de l'insertion ou de la délétion de séquences répétées à l'intérieur des exons, s'ils commencent à être élucidés, restent néanmoins encore assez mal connus. Nous l'avons vu, le système de réplication de l'ADN se compose d'un complexe enzymatique capable de détecter et de corriger les erreurs de réplication. Cependant, la nature du brin matrice à dupliquer a une grande influence sur la capacité de ce complexe enzymatique à assurer la fidélité, la présence de séquences répétées semblant favoriser les erreurs d'insertions et/ou de délétions de bases. Les mécanismes d'apparition et de duplication de séquences répétées ont été essentiellement étudiés chez les microsatellites qui sont des séquences d'ADN composées d'un nombre variable de répétitions en tandem (c'est-à-dire insérées dans le même sens, à l'inverse des séquences répétées inversées) et servant de marqueur moléculaire d'espèce. Il a ainsi été déterminé que les répétitions seraient essentiellement dues à des dérapages de la polymérase, une sorte de bégaiement. En effet, lors de la progression de l'enzyme le long d'une séquence répétée, le brin matrice et le brin néo-synthétisé peuvent se ré-apparier mais dans une configuration non complémentaire, provoquant donc un mésappariement. Ainsi, si l'élément répété mésapparié se situe désormais sur le brin matrice, le brin fille s'allongera d'un élément (il va sans dire que dans le cas inverse, le brin fille sera amputé de cet élément) (Sia *et al.* 1997 ; Umar & Kunkel 1996). Ainsi, au fur et à mesure des réplications successives, un nombre croissant de séquences répétées peut être intégré, ou retiré, dans les brins d'ADN néo-synthétisés.

Introduisons ici le concept d'homologie. *Stricto sensu*, l'homologie fait référence au fait que les similarités que l'on peut observer entre deux séquences sont issues de leur évolution à partir d'un ancêtre commun. Cette notion remonte initialement au XIXème siècle où elle fait alors référence à des structures dont l'organisation fondamentale serait conservée entre des organismes différents (Owen 1848). Jusqu'alors, nous l'avons vu, les systèmes de classifications développés au cours des siècles ne tenaient compte que de l'observation macro-phénotypique des organismes étudiés, c'est-à-dire de leurs homologies morphologiques, anatomiques ou comportementales. La phylogénèse des organismes se heurtait à un écueil insurmontable : l'absence d'outils moléculaires permettant l'étude des homologies entre organismes phénotypiquement très éloignés ou au contraire indifférentiables. C'est d'ailleurs la raison pour laquelle le monde des microorganismes a été si longtemps tenu à l'écart de toute étude phylogénétique. Si la découverte de l'ADN par Watson et Crick correspond à notre sens à la seconde révolution conceptuelle après celle de Darwin, l'étude des informations contenues dans cette molécule n'était réalisable qu'à une

très petite échelle, et donc son utilisation dans des études phylogénétiques restait marginale. Cependant, l'accumulation de données, même partielles, sur l'ADN ouvrit la voie à un concept nouveau, celui que Woese et Fox, en 1977, appelèrent LUCA (Woese & Fox 1977). LUCA, pour « *Last Unique Common Ancestor* », est la séquence ancestrale à partir de laquelle toutes les autres séquences ont dérivé. Il représente donc le précurseur unique à l'origine de toute l'évolution du vivant. Emergé il y a plusieurs milliards d'années, il a donné lieu par spéciations successives, aux trois règnes qui désormais ont été identifiés : les eucaryotes (organismes possédant un noyau), et les procaryotes (organismes sans noyau) eux-mêmes séparés en deux règnes, les eubactéries et les archées. Les archées se distinguent des eubactéries par leur membrane plus résistante et un système de réplication de leur génome plus proche des eucaryotes. La dichotomie du vivant en deux règnes, animal/végétal, puis eucaryotes/procaryotes qui avait prévalu jusqu'alors venait de s'éteindre. Le dogme des trois règnes du vivant n'a depuis jamais été remis en cause. Reste l'énigme des populations de virus, que l'on peut difficilement classer dans aucun des trois règnes. L'on s'interroge même de savoir si un virus peut, ou non, être considéré comme un être vivant. A notre sens, les virus peuvent représenter la frontière entre le vivant et l'inerte.

Une troisième révolution dans ces approches sur la diversité, eut lieu en 1983. Cette révolution ne fut, cette fois, pas conceptuelle, mais technique : la possibilité d'amplifier exponentiellement *in vitro* n'importe quel fragment d'ADN à partir d'une enzyme ADN-polymérase-ADN-dépendante et de deux courtes amorces d'ADN permettant l'initiation de la polymérisation. Cette technique de polymérisation en chaine (PCR pour « *polymerase chain reaction* ») fut mise au point par K. Mullis (Mullis 1986 ; Mullis & Faloona 1987). Cette découverte autorisait désormais l'étude à grande échelle du support de l'information génétique, donc la compréhension des mécanismes de la vie, et ouvrait la voie à la phylogénèse des organismes sur le plan moléculaire. Ce pas de géant dans la capacité à comprendre le fonctionnement du vivant valut à Kary Mullis le prix Nobel de chimie en 1993.

1-3-1-2- Les recombinaisons

Comme nous venons de le voir dans le chapitre précédent, l'évolution des séquences d'ADN peut être spontanée, par interversion de nucléotides ou erreurs de la polymérase lors de la réplication du brin d'ADN matriciel. Cependant, d'autres mécanismes existent permettant l'évolution des organismes, mécanismes cette fois non réellement stochastiques, car mettant en œuvre toute une batterie d'enzymes telles que les endonucléases et les ligases. L'un de ces mécanismes est la recombinaison.

Le phénomène de recombinaison apparaît au cours de la réplication de l'ADN, il se caractérise par l'échange de fragments d'ADN homologues entre deux génomes différents, permettant de créer des génomes nouveaux, et donc de participer activement au maintien

de la diversité à l'intérieur des populations. Notons que, si la plupart du temps l'échange d'information génétique est à base d'ADN, il peut aussi avoir pour base de l'ARN, notamment chez certains virus, tels le virus de la maladie de Newcastle (NDV) ou de la grippe aviaire (AIV). Certains virus sont, en effet, les seuls organismes dont l'information génétique peut ne pas être portée par de l'ADN mais par de l'ARN (il existe 15 familles de virus à ARN). Selon le règne auquel appartiennent les espèces, le phénomène de recombinaison s'effectue à des moments privilégiés spécifiques.

Chez les eucaryotes, c'est principalement au cours de la méiose, c'est-à-dire de la reproduction sexuée, que les recombinaisons se réalisent. Les eucaryotes sont des êtres diploïdes, c'est-à-dire qu'ils possèdent une paire de chaque chromosome, donc de chaque gène qu'ils portent. Durant la formation des gamètes, les paires de chromosomes se mêlent et s'entrecroisent, permettant à des loci homologues d'être échangés d'un chromosome vers un autre (Borde 2007).

Chez les procaryotes, plusieurs mécanismes permettent l'apparition de recombinaisons entre les génomes d'organismes différents. Il s'agit de la conjugaison, qui est l'échange entre deux bactéries de plasmide(s), molécules d'ADN circulaires contenant de l'information génétique. Ces plasmides peuvent rester en position épisomale, c'est-à-dire libre dans le cytosol bactérien, ou bien s'intégrer dans génome de la bactérie receveuse (Clewell 2011). C'est une méthode dite sexuée. Le second est la transduction, mécanisme au cours duquel l'information génétique est transmise d'une bactérie donneuse vers une bactérie receveuse non pas par l'intermédiaire d'un plasmide, mais par le biais d'un vecteur viral, ou bactériophage (Miller 2001). Enfin, il existe un processus de transformation, durant lequel les bactéries acquièrent et incorporent des gènes provenant de leur environnement.

Les virus, quant à eux, ne pratiquent pas la reproduction. En effet, même si certains d'entre eux codent pour des polymérases qui leur sont propres (tel le virus de la peste porcine africaine), ils ne disposent pas des complexes enzymatiques leur permettant de se répliquer, ni d'assurer la traduction de leurs gènes de façon autonome. Ce sont donc des parasites stricts obligatoires qui, après avoir infecté une cellule, en détourne la machinerie dans le seul but de leur réplication. La recombinaison de l'ADN ou de l'ARN de ces organismes ne devrait donc avoir lieu que lorsque deux virus infectent simultanément une cellule. C'est le cas, par exemple, avec le virus de la grippe aviaire, un virus à ARN segmenté, pour lequel l'échange de fragments entre différentes souches virales donne lieu à l'apparition de nouveaux variants. On parle alors de virus réassortis. Il est à noter que le réassortiment peut également avoir lieu entre des virus proches, infectant le(s) même(s) hôte(s), mais appartenant à des familles virales différentes. Ces recombinaisons peuvent alors donner lieu à l'émergence de nouveaux genres viraux, et même à de nouvelles maladies (Garcia-Arenal et al. 2001). Cela a été observé, notamment chez les virus à ARN, entre les virus GFLV (*Grapevine fanleaf virus*) et ArMV (*Arabis mosaic virus*), qui peuvent co-infecter la vigne (Vigne et al. 2008). La réplication des virus se fait donc à partir des

complexes enzymatiques cellulaires. Or, si une cellule duplique son ADN dans le but de se diviser, le génome viral, lui, est répliqué en continu, dans le but de produire un maximum de nouveaux virions. Dans les « usines à virus », qu'elles soient nucléaires ou périnucléaires, se trouve donc un nombre très élevé de copies de génomes viraux et de fragments de génomes qui se mêlent et s'entrecroisent avant d'être assemblés. Les recombinaisons qui s'opèrent alors tiennent du même mécanisme que celui qui a lieu lors de la méiose des cellules eucaryotes. Notons que la plupart de ces recombinaisons seront létales pour le virus, le génome viral pouvant être amené à varier de façon trop importante.

1-3-1-3- Les transferts latéraux de gènes

Le transfert latéral de gènes consiste pour un organisme, à recevoir de l'information génétique ne provenant pas de son ou de ses ascendants. Il a été décrit pour la première fois en 1960 par l'observation du passage de gènes de résistance aux antibiotiques entre des espèces de bactéries différentes (Akiba *et al.* 1960). Ce mécanisme ne semble pas aujourd'hui présenter un grand intérêt pour les organismes eucaryotes pluricellulaires en termes d'évolution, mais il a sans doute joué, et joue toujours aujourd'hui, un grand rôle dans la diversité des organismes unicellulaires, qu'ils soient procaryotes ou eucaryotes (Bock & Timmis 2008).

Les virus Les virus sont eux aussi sujets aux transferts latéraux de gènes. Des travaux récents ont émis l'hypothèse que les grands virus à ADN nucléocytoplasmiques (NCLDVs, pour *nucleocytoplasmic large DNA viruses*, famille dont fait partie le virus de la PPA), et dont le génome varie entre 100 kb et 1,2 Mb et code pour de très nombreuses protéines, pouvait, à l'origine, n'être composé que de quelques gènes. Au cours de leur évolution, ils auraient ainsi acquis par transfert latéral de nombreux gènes provenant des cellules des hôtes qu'ils infectent (Filee & Chandler 2010). On a également découvert que ces virus qualifiés de « géants » pouvaient eux-mêmes être infectés par des virus plus petits qui non seulement détournent la machinerie virale mise en place dans les usines à virus, mais capterait également certains de leurs gènes pour les intégrer dans leur propre génome. C'est le cas, par exemple, du mimivirus *Acanthamoeba polyphaga*, infecté par un virus du genre mamavirus appelé « Sputnick » (La Scola 2008).

1-3-1-4- L'Hybridation

L'hybridation est enfin le dernier mécanisme permettant l'évolution des organismes. Elle consiste en un croisement de deux espèces différentes afin d'en créer une nouvelle. Ce

mécanisme ne concernant pas les virus, dont la multiplication est asexuée, nous ne le développerons pas ici.

1-3-2- Les processus de l'évolution des séquences d'ADN : les forces évolutives

1-3-2-1- La dérive génétique

Littéralement, la dérive génétique correspond à la fluctuation stochastique de la fréquence des allèles provoquée par l'échantillonnage aléatoire des gamètes à l'origine des nouveaux individus dans une population donnée (Masel 2011). Ce concept n'intervient pas *sensu stricto* dans les mécanismes de la diversité virale, puisque la multiplication des virus n'a pas de base somatique, mais *sensu lato* par le fait que la fluctuation due à l'échantillonnage des organismes est à la base du principe de la théorie neutraliste. Deux mécanismes entrainent ainsi principalement la dérive génétique des virus, influençant grandement la diversité virale : la cassure antigénique (*antigenic shift*) et la dérive antigénique (*antigenic drift*).

La notion de cassure antigénique est essentiellement employée pour caractériser l'évolution des virus *influenza* de la grippe. En effet, elle est le fruit d'un processus au cours duquel plusieurs isolats viraux de la même espèce, ou d'espèces très semblables, se recombinent pour former un virus mosaïque des virus originaux et dont les antigènes de surface vont être également composites, représentant un nouveau phénotype. La recombinaison peut porter sur l'intégralité d'un segment (échange de segment ou réassortiment) ou se limiter à des échanges intra-segment (recombinaison *sensu stricto*). La variation phénotypique induite sur les antigènes permettra au virus recombiné de ne plus être reconnu par les anticorps du système immunitaire de l'hôte. Le réassortiment est comparable aux recombinaisons que nous avons évoquées plus haut et apparaît lors d'une co-infection. Il est de plus particulièrement fréquent chez les virus capables d'infecter une gamme variée d'hôtes, dont le virus de la grippe est un parfait exemple, puisqu'il peut infecter les oiseaux, les porcs, et les humains (Treanor 2004). Ce mécanisme est ainsi à l'origine de l'émergence de nombreux nouveaux virus.

La dérive antigénique résulte de l'accumulation de mutations stochastiques naturelles ayant lieu dans les séquences au cours du temps, mutations entrainant des modifications phénotypiques des antigènes viraux, et pouvant aboutir à un échappement à la réponse immune acquise de l'hôte. La dérive antigénique peut être plus ou moins importante selon la durée de l'infection et la réponse immunitaire de l'hôte. En effet, une infection perdurant dans le temps entrainera une réponse plus importante du système immunitaire, et favorisera ainsi la dérive antigénique. Les variants étant moins sujets à la reconnaissance par

les anticorps du système immunitaire, ils seront alors favorisés tant pour leur réplication chez l'hôte que pour leur dissémination.

Ces deux mécanismes participent de la sélection naturelle des variants biologiques, sélection qui peut prendre plusieurs formes.

1-3-2-2- La pluralité des sélections naturelles

Comme nous l'avons vu lors de la description de la théorie de l'évolution de Darwin, les forces naturelles tendraient à sélectionner les organismes les plus adaptés à leur environnement, et à favoriser le maintien et la propagation de leurs caractères par le truchement de leur(s) descendance(s). Deux conceptions, néanmoins, s'affrontent à ce sujet : la théorie *sélectionniste* et la théorie *neutraliste*.

La théorie sélectionniste affirme que les mutations nucléotidiques acquises de façon aléatoire, pour pouvoir être conservées au cours de l'évolution, doivent conférer un avantage additionnel pour la survie de l'organisme ou de son espèce. Il n'est pas question ici de supériorité, mais d'adaptation spécifique au milieu. Ainsi, l'hypothèse d'un milieu confiné, totalement indépendant, devrait permettre l'émergence d'espèces dont le phénotype diffèrerait d'espèces analogues, mais ayant prospéré dans un environnement différent, ce que l'on appelle des niches écologiques. Cette hypothèse est validée lorsque l'on étudie l'évolution d'êtres vivants ayant été séparés durant des millénaires de leur ancêtre originel. C'est le cas par exemple de Madagascar ou encore des îles Galápagos dans lesquelles Darwin fit de longues observations de la faune et de la flore, îles où l'homme n'avait jamais exercé de sélection artificielle des êtres vivants. L'évolution d'organismes en fonction d'un environnement géographique donné est appelée *spéciation allopatrique* (Bolnick 2007). Il semble par ailleurs évident que les protéines codées par exemple par une bactérie halophile aient des caractéristiques physico-chimiques différentes de celles codées par une bactérie thermophile (Fukuchi 2003 ; Lobry 2006). Les contraintes thermodynamiques environnementales semblent donc jouer un grand rôle dans la structure primaire des protéines et ainsi dans la composition des codons des gènes des organismes exposés à ces contraintes.

Plusieurs types de sélections sont à l'œuvre pour guider l'évolution des organismes. En premier lieu, on trouve la sélection *stabilisante*. Cette forme de sélection naturelle tend à éliminer les caractéristiques phénotypiques extrêmes des organismes lorsqu'elles leur sont défavorables dans leur environnement. Elle favorisera donc une normalisation des phénotypes, avec pour effet induit de réduire la diversité au sein d'une population. A l'inverse, on observe une sélection dite *diversifiante* ou *disruptive*. Cette sélection naturelle tend à favoriser les extrêmes phénotypiques aux dépends des phénotypes normaux ou moyens qui sont les plus représentés. Ce type de sélection, accentuant la divergence entre

individus, peut aisément mener à la spéciation des populations. Il indique que des conditions différentes peuvent être tout aussi favorables à l'établissement et à la survie de phénotypes très différents, voire opposés. Cette séléction a de plus l'avantage d'offrir de plus grandes chances de préserver l'espèce en cas de bouleversement majeur du contexte environnemental puisqu'elle permet de maintenir une diversité de phénotypes dont certains seront plus favorables à l'adaptation à un nouvel environnement. Ces deux types de sélection naturelle sont dits directionnels, car ils tendent à favoriser un ou des phénotypes à l'intérieur d'une population donnée.

Un troisième type de sélection est à l'œuvre dans l'évolution. Il s'agit de la sélection *équilibrée*. Dans ce processus sélectif, les différents allèles d'un même gène vont être maintenus au sein d'une population, avec une fréquence supérieure à celle des mutations naturelles pouvant apparaître sur ce gène. Cette forme de sélection naturelle est principalement réalisée au sein de populations dont les individus sont diploïdes, l'hétérozygotie apportant un avantage par rapport à l'homozygotie (King R.C. 2006). Ce n'est évidemment pas le cas pour les populations virales. Ce mode de sélection favorisera et maintiendra un haut niveau de diversité au sein des populations qu'on appelle panmictiques, c'est-à-dire où tous les individus ont la même probabilité de transmettre leurs gènes à la descendance.

De la même manière qu'évolutionnistes *fixistes* et *transformistes* se sont affrontés au cours du XIXème siècle, une controverse existe également entre les tenants de l'évolution *sélectionniste* et de l'évolution *neutraliste*.

En biologie de l'évolution, un caractère *neutre* ne présente ni avantage, ni désavantage pour l'organisme qui le porte. Il n'est donc pas soumis à la sélection naturelle. L'évolution d'un tel caractère est totalement indépendante de son environnement et la dérive génétique n'a aucune influence sur lui. Bien évidemment, une telle indépendance ne peut être absolue, et doit donc être appréhendée de façon relative. On dira donc que la pression de sélection qui l'affecte est plus ou moins forte et plus ou moins directionnelle. Ce caractère sera donc plutôt soumis à une évolution plus prosaïquement stochastique.

La théorie neutraliste s'applique principalement aux séquences d'ADN et de protéines plus qu'aux organismes entiers. Elle a été formulée pour la première fois par Motoo Kimura (1924 – 1994), vers la fin des années 1960 (Kimura 1968). Elle postule que la majorité des substitutions observées chez les variants ne sont pas le résultat d'une sélection darwinienne (sélection naturelle positive), mais le fruit d'une fixation stochastique de mutations neutres. Pour étayer sa théorie, Kimura s'est basé sur un constat simple, celui que la majorité des substitutions nucléotidiques intervient sur la position 3 des codons, et que 72% des mutations sur cette position sont silencieuses, c'est-à-dire qu'elles n'entrainent pas la mutation de l'acide aminé dans la protéine traduite. Cela revient à dire que la majorité des mutations ayant lieu dans la séquence nucléotidique des gènes ne confèrent aucun avantage

évolutif à l'organisme qui les reçoit et par la même aucune adaptation bénéfique à son environnement (Kimura 1977).

La conséquence principale de cette théorie, à l'inverse de la sélection naturelle, est qu'elle ne mène pas systématiquement à la complexification des organismes. Chez les organismes les plus simples tels que les procaryotes, l'évolution conduit non pas à une complexification mais à une diversification de plus en plus importante. C'est ce que Stephen Jay Gould nomme « le mur de gauche », c'est-à-dire la tendance statistique naturelle de tout organisme acculé contre un mur à vouloir l'éviter. Pour illustrer son propos, Gould prend l'exemple des bactéries qui ont colonisés toutes les niches écologiques. C'est l'absence d'évolution vers forcément davantage de complexification qui provoqua la polémique avec les tenants du sélectionnisme darwinien. Cependant, il est possible de ne pas opposer radicalement ces deux théories mais au contraire d'en faire des mécanismes complémentaires : ainsi, la sélection adaptative garde une place prépondérante dans l'évolution mais sans doute pas davantage que l'évolution purement stochastique des séquences d'acide nucléique.

L'avantage principal de la théorie neutraliste est qu'elle parvient à réunir la génétique mendélienne établissant la transmission de l'hérédité, la théorie darwinienne de la sélection naturelle ainsi que la génétique des populations introduite par Fisher. La réunion de ces trois concepts est nommée « *théorie synthétique de l'évolution* » (Dobzhansky 1937 ; Haldane 1932 ; Huxley 1942). Comme toute théorie basée sur des modèles mathématiques, la théorie neutraliste repose sur des postulats et ce sont ces hypothèses qui font sa faiblesse. En effet, elle présuppose que (i) le taux de substitution des nucléotides est constant dans le temps (hypothèse de l'horloge moléculaire stricte), ce qui est très rarement vérifié, (ii) que la taille des populations est également constante dans le temps, ce qui est, là aussi, rarement le cas, surtout lors des périodes d'extinction ou de spéciation massives, et (iii) qu'il existe un équilibre entre les allèles créés et disparus lors de la dérive génétique, ce qui reste à démontrer.

2- Les méthodes d'analyse

2-1- La phylogénie moléculaire

La phylogénie moléculaire utilise pour les comparer les séquences des molécules d'ADN ou de protéines des êtres vivants dans le but de déterminer les liens de parenté qui les unissent ainsi que pour appréhender leur histoire évolutive (phylogénèse). Suite à la découverte par Watson et Crick de la structure de la molécule d'ADN, de nombreuses techniques ont été mises au point pour décoder l'enchainement des quatre nucléotides qui la constitue. L'avènement de la PCR et du séquençage par la technique de Sanger (Sanger 1977) a permis de générer des données de plus en plus massives de séquences d'ADN,

données qui ont permis le développement de l'approche phylogénétique pour la classification des êtres vivants. L'analyse en phylogénie moléculaire, qui prend en compte les caractéristiques physico-chimiques des molécules étudiées, génère des arbres phylogénétiques.

Un arbre phylogénétique schématise donc les liens unissant les taxons (ou organismes). Les feuilles de l'arbre sont représentées par les séquences connues et analysées des taxons, tandis que les nœuds intermédiaires figurent les ancêtres à l'origine de ces taxons et dont les séquences sont inconnues. C'est la raison pour laquelle on parle d'inférence pour qualifier la manière de calculer un arbre phylogénétique car il forme la conclusion d'interprétations faites par des modèles mathématiques à partir de données réelles. Les taxons se situant après un nœud intermédiaire dans l'arbre, taxon hypothétique dont ils sont les descendants, forment des clades. Enfin, un arbre peut être enraciné si l'ancêtre commun de toutes les séquences étudiées a pu être déterminé préalablement.

La cladistique est une théorie qui a été développée dans les années 1950 par l'entomologiste allemand Willi Henning (Henning 1950). Elle a jeté les bases sémantiques permettant de nommer phénotypiquement les taxons, les groupes de taxons, ainsi que les caractères permettant leur comparaison. Ainsi, un caractère dans une séquence est dit *apomorphe* lorsque, quoique différent du caractère ancestral, il est cependant pertinent en termes de différenciation entre deux organismes semblables. Si ce caractère, malgré une ascendance commune, est unique pour un taxon considéré, on dira alors de lui qu'il est *autopomorphe*. *A contrario*, des états de caractères peuvent être partagés entre des taxons, sans pour autant permettre de les regrouper en clades. Ces caractères sont dits *plésiomorphes*. Enfin, la *synapomorphie* décrit des caractères partagés par deux descendants au moins et suffisamment pertinents en termes d'informations phylogénétiques pour reconstruire l'histoire évolutive de ce groupe d'organismes. Le choix des données qui vont être étudiées est donc crucial pour que les résultats de l'analyse soient pertinents, et retracent au plus près la véritable histoire évolutive des séquences et donc des organismes qui les portent.

2-2- Le choix des données

Les relations entre deux organismes ont tout d'abord été établies par comparaison de leurs caractères morphologiques. Depuis la recrudescence des données moléculaires, c'est-à-dire depuis les années 1990, la comparaison des séquences nucléiques et protéiques pour expliquer les liens entre organismes a largement supplanté la comparaison morphologique. Néanmoins, pour les espèces éteintes, dont il ne reste que des fossiles, cette dernière méthode reste encore la seule envisageable. Concernant les virus, il n'existe pas de fossile connu à ce jour. La seule façon d'étudier leurs relations est donc restreinte à l'utilisation de

séquences provenant de virus contemporains. Littéralement, les méthodes phylogénétiques étudient les similarités entre les séquences géniques en partant de l'hypothèse qu'elles sont homologues et donc partagent un ancêtre commun, quel que soit son âge. Cependant, la durée de l'histoire évolutive d'un gène peut avoir engendré une séquence si différente de l'originale que les informations qu'elle porte ne sont plus suffisantes pour que le résultat de leur comparaison soit encore probant. Dans le cas de gènes ayant trop largement divergés, on ne parle d'ailleurs plus d'homologie, mais de similarité, le terme d'homologie étant désormais réservé à des séquences ayant un ancêtre commun récent.

L'ADN étant composé de seulement quatre bases différentes, la comparaison de deux séquences choisies aléatoirement, c'est-à-dire non homologues, montrera en moyenne 25% de nucléotides identiques. L'arbre phylogénétique inféré d'une telle comparaison verrait alors sa valeur prédictive considérablement amoindrie, pour ne pas dire inexploitable. On considère donc que pour comparer des séquences nucléotidiques, il convient qu'elles partagent au minimum 60% de leurs caractères. Les séquences comparées seront donc très proches, ne divergeant que par quelques points de mutations, et partageront une histoire évolutive commune. Nous l'avons vu, les différences entre les séquences peuvent avoir diverses origines et la condition d'homologie n'est pas nécessairement suffisante pour retracer l'histoire évolutive d'un gène. Si les divergences constatées au sein d'une séquence génique proviennent d'une duplication, c'est-à-dire d'une copie supplémentaire d'un gène au sein d'un même génome, ces gènes seront dit paralogues (Fitch 1970). Leur divergence, indépendante l'une de l'autre, sera sans lien avec l'évolution de l'organisme qui les porte. En revanche, si l'évolution de deux gènes homologues a eu lieu après la spéciation et a donc été indépendante, ils seront dits orthologues. Le choix des séquences à comparer dépendra donc du résultat recherché. Ainsi, l'utilisation de gènes orthologues permettra de s'intéresser à la spéciation, tandis que celle de gènes paralogues étudiera leur duplication, c'est-à-dire leur évolution au sein d'un même organisme. Notons que le phénomène de duplication semble avoir joué un rôle important dans l'évolution des espèces, en leur permettant d'acquérir de nouvelles fonctions (Ohno 1970). Outre la duplication d'un gène au sein d'un même génome pouvant biaiser la phylogénèse d'une espèce par le phénomène de *paralogie cachée*, le transfert latéral de gènes entre deux espèces différentes peut également amener à rapprocher deux espèces pourtant éloignées dans l'arbre du vivant. Il conviendra donc, pour mener à bien une étude phylogénétique, de non seulement analyser les arbres obtenus, mais de croiser ces résultats avec l'histoire naturelle des organismes étudiés.

L'analyse phylogénétique portant sur les gènes homologues, ne devrait cependant pas se limiter à une étude globale. En effet, selon les forces évolutives à l'œuvre et la structure même des séquences, qu'elles soient d'ADN ou d'acides aminés, des parties des séquences peuvent évoluer de façon différente. Ainsi, par exemple, les sites catalytiques d'une enzyme auront nettement moins tendance à varier afin que leur fonction puisse être préservée. Plus que l'information contenue dans les séquences homologues, ce sont donc les sites

homologues présents dans ces séquences qui vont faire l'objet de toute notre attention. Or, il existe une condition *sine qua non* permettant l'analyse de tels sites, c'est que leur alignement soit correct entre les séquences d'intérêts, c'est-à-dire qu'il reflète la réalité structurelle de ces séquences. Pour cela, les sites homologues doivent être alignés sous forme de colonnes.

Il existe aujourd'hui de nombreux algorithmes permettant d'aligner des séquences homologues. Initialement, les alignements furent effectués manuellement. Ils prenaient la forme de matrice de points dont les colonnes représentaient les sites homologues et les lignes les séquences. Cette méthode « ancestrale », nonobstant son côté fastidieux, avait le désavantage de situer avec difficulté les indels dans les séquences, et donc ne pouvait garantir que l'alignement généré fut optimal. Le premier algorithme permettant une automatisation des alignements de séquences fut développé dès 1970 par Needleman et Wunsh (Needleman 1970). Il utilisait une programmation dynamique, dans laquelle un score était établi selon un indice de similarité entre deux séquences, l'algorithme maximisant ce score pour établir un alignement optimal. A la même période, l'algorithme que développa Sellers considérait, lui, la distance entre deux séquences pour la minimiser au maximum (Sellers 1974).

Si ces algorithmes peuvent en théorie être applicables pour l'alignement de plus de deux séquences, il s'avère que le temps de calcul et la capacité informatique nécessaires rendent leur utilisation en pratique impossible. Des algorithmes d'alignements multiples ont alors été développés qui fournissent une approximation de l'alignement optimal des séquences mais permettent l'alignement de très nombreuses séquences, au moyen d'alignements progressifs. Cette méthode commence par aligner les séquences les plus semblables, puis les autres séquences sont implémentées successivement dans l'alignement de la plus similaire à la plus dissemblable (Corpet 1988 ; Taylor 1987). Parmi les algorithmes les plus récents et les plus utilisés, on pourra citer l'alignement multiple par les méthodes ClustalW (Higgins *et al.* 1996) et MUSCLE (Edgar 2004). Il est important de noter ici que, si les algorithmes permettant de réaliser des alignements multiples de séquences d'ADN sont parfaitement à même de détecter les zones d'homologies entre les séquences, ils ne peuvent pas déterminer si l'alignement produit tient ou non compte du cadre de lecture des gènes. Dans le cas où les séquences analysées sont codantes, c'est d'un intérêt fondamental pour respecter au mieux la réalité biologique des sites alignés. Tout alignement généré par ces algorithmes doit donc être contrôlé minutieusement avant reconstruction phylogénétique. Les alignements ainsi vérifiés vont être à la base de toute analyse phylogénétique, puisqu'ils mettent en exergue les divergences accumulées entre les sites homologues des séquences au cours de leur histoire évolutive.

2-3- Reconstructions phylogénétiques

Il existe plusieurs méthodes de reconstruction permettant d'inférer un arbre phylogénétique. La première est basée sur l'observation simple de la distance entre des séquences prises deux à deux, établie en termes de différences en nucléotides ou en acides aminés. La méthode des distances est représentée par la méthode UPGMA (tombée en désuétude) ainsi que la méthode BioNJ. Ces méthodes construisent des arbres non enracinés en incrémentant les séquences en fonction de leur proximité avec les précédentes. Elles suivent un modèle évolutif explicite, c'est-à-dire basé sur l'observation des mécanismes évolutifs à l'œuvre. La seconde s'intéresse à l'état des caractères qui divergent à chaque site des séquences (les colonnes de l'alignement), c'est-à-dire qu'elle prend en compte la nature et la position des substitutions (transversion, transition, indel). C'est le cas des méthodes de maximum de parcimonie, de maximum de vraisemblance et d'inférence bayésienne. Ces dernières peuvent être subdivisées en deux groupes : le maximum de vraisemblance et l'inférence bayésienne, qui construisent un arbre phylogénétique en suivant un modèle explicite d'évolution, tout en calculant la probabilité de l'organisation des branches de l'arbre, ainsi que de la longueur de ces branches. Ces méthodes génèrent des arbres enracinés puisque les modèles appliqués tiennent compte de la flèche du temps. Le maximum de parcimonie, quant à lui, ne suit pas de modèle explicite d'évolution, puisqu'il se borne à comptabiliser le nombre de « pas mutationnels » nécessaires pour passer d'une séquence à une autre au sein de l'arbre.

Dans les deux cas, ces méthodes d'inférence phylogénétique s'appuient sur une matrice de distance qui est calculée à partir de l'alignement des séquences étudiées. La question fondamentale sous-jacente à la génération de ces matrices est la suivante : comment peut-on définir mathématiquement la distance évolutive, ou distance phénétique, entre deux séquences qui s'approchent au plus près de la réalité biologique de cette divergence. La nature stochastique des substitutions apparaissant dans des séquences au cours du temps est un concept assez partagé. Il convient donc d'établir un modèle de ce processus stochastique d'évènements de substitution.

La méthode la plus simple d'accès pour évaluer la distance entre deux séquences consiste à calculer la proportion de sites homologues divergents. Cette mesure est appelée *p-distance* et elle est exprimée en nombre de substitutions par site existant entre deux séquences. Si elle est simple à effectuer, cette méthode ne peut rendre compte de la distance phénétique réelle entre deux séquences. D'une part, elle ne prend pas en compte la possibilité des substitutions multiples sur un même site (exemple, la substitution observée d'une Thymine par une Guanine mais qui s'est déroulée par l'intermédiaire, non observable d'une Thymine en Cytosine puis en Guanine). D'autre part, elle ne peut non plus prendre en compte les phénomènes de réversion calculant une *p-distance* égale à 0 entre deux sites qui ont pourtant eu une histoire évolutive différente. Le calcul de la distance phénétique observée revient donc le plus souvent à sous-estimer la distance génétique réelle entre des

séquences homologues. C'est la raison pour laquelle il convient de modéliser l'évolution afin de pouvoir corriger la mesure de la *p-distance*.

Plusieurs formules permettant de calculer la distance entre des séquences ont été formulées au cours du temps. Parmi elles, on trouvera celle énoncée par Nei en 1972 (Nei 1972) et en 1978 (Nei 1978) ou par Reynolds en 1983 (Reynolds *et al.* 1983).

Le principe des méthodes de modélisation de l'évolution par les distances tentent de générer un arbre phylogénétique à partir d'une matrice de distances établissant la distance génétique séparant des séquences deux à deux (Felsenstein 1988). Or, comme nous venons de le voir, la *p-distance* est une sous-estimation de la distance génétique réelle, et c'est pourquoi l'on va chercher à appliquer un modèle évolutif le plus réaliste possible pour les séquences considérées. Notons que l'utilisation d'un modèle irréaliste entrainera d'énormes biais dans la détermination de la topologie de l'arbre inféré à partir de la matrice (Lockart 1994 ; Van de Peer 1996).

On distingue donc les méthodes de reconstruction selon qu'elles soient basées sur les caractères ou non et sur un modèle d'évolution ou non (Tableau 1).

Méthodes	**Basée sur les caractères**	**Non basée sur les caractères**
Basée sur un modèle explicite d'évolution	1-Maximum de vraisemblance 2-inférence bayésienne	Distance (BioNJ, UPGMA)
Non basée sur un modèle explicite d'évolution	Maximum de parcimonie	

Tableau 1 : caractéristiques des différentes méthodes de reconstruction phylogénétique. Les méthodes basées sur les caractères regardent leur état à chaque site de la séquence tandis que les méthodes basées sur les distances s'intéressent à la proximité entre les séquences.

2-4- Reconstructions phylogénétiques par la méthode des distances : UPGMA, minimum d'évolution et méthode du plus proche voisin

La première méthode développée fût l'UPGMA (pour *Unweighted – Pair Group Method with Arithmetic means*). Elle est aujourd'hui tombée en désuétude, car une des hypothèses sur lesquelles elle repose est l'hypothèse de l'horloge moléculaire stricte, c'est-à-dire un taux de substitution constant dans toutes les branches de l'arbre, arbre dit alors ultramétrique. Comme cela n'arrive pour ainsi dire jamais, cette méthode est donc extrêmement sensible à un taux de substitution variable selon les phylum (Huelsenbeck

1993). De plus, un arbre ultramétrique doit être enraciné, c'est-à-dire contenir un groupe externe, et dans lequel tous les taxa sont équidistants de la racine, condition qui n'est pas toujours réalisable.

D'autres algorithmes, comportant moins de biais analytiques ont été alors développés : la méthode du « minimum d'évolution » (ME) (Kidd 1971 ; Rzhetsky 1992b) et la méthode du plus proche voisin, ou « Neighbor-Joining » (NJ) (Saitou & Nei 1987). La méthode ME propose d'examiner toutes les topologies des arbres possibles et d'en calculer la longueur totale des branches S. La topologie retenue étant celle pour laquelle la longueur S est minimale. L'une des limitations de cette méthode est qu'elle se veut heuristique, c'est-à-dire analysant successivement toutes les hypothèses possibles. La méthode NJ, pour sa part, est une approximation du ME. Bien que reposant également sur une heuristique comme le ME, il a été montré que les arbres produits par cette méthode était cependant très similaires à ceux générés par le ME (Pauplin 2000 ; Rzhetsky 1992a).

Quand bien même ces méthodes de reconstruction produisent des arbres phylogénétiques fiables, elles sont soumises au phénomène dit d'*attraction des longues branches*, biais d'analyse qui aura tendance à considérer des séquences très divergentes comme des séquences sœurs et donc à les regrouper au sein d'un même clade.

2-5- Reconstruction phylogénétique par le maximum de parcimonie

L'analyse phylogénétique par le maximum de parcimonie (MP) met en œuvre un critère permettant d'estimer et de minimiser le nombre d'évènements évolutifs ayant permis le passage d'une séquence à une autre. Autrement dit, elle recherche le plus petit nombre de changement d'état des caractères composant les séquences. Cette méthode a tout d'abord été développée pour la comparaison de données morphologiques (Hennig 1966). Elle tire son origine d'un concept philosophique, dit du *rasoir d'Ockham*. Guillaume d'Ockham fut un philosophe franciscain (1285 – 1347) rationaliste qui postulat le concept suivant : « *Pluralitas non est ponenda sine necessitate* » (les multiples ne doivent pas être utilisés sans nécessité), ce qui signifie que les hypothèses les plus simples sont souvent les plus vraisemblables (nous verrons ultérieurement que cette assertion doit aussi être utilisée lors du choix d'un modèle évolutif, celui prenant en compte le moins de paramètres devant être privilégié). C'est pourquoi ce principe de parcimonie est également appelé principe de simplicité ou d'économie.

A la différence des méthodes basées sur les distances entre les séquences, l'approche par le maximum de parcimonie tient compte de l'état individuel de chaque caractère contenu dans les séquences. Elle a pour principales hypothèses :

- l'indépendance des sites, c'est-à-dire que les évènements affectant un caractère précis de la séquence n'influencent pas ni ne sont influencés par les évènements affectant les autres caractères de la séquence (ce qui n'est pas toujours vérifié, les structures secondaires des caractères pouvant avoir une influence sur leurs voisins (Tillier & Collins 1998)),

- l'uniformité du processus d'évolution, c'est-à-dire l'homogénéité du taux d'évolution pour tous les sites d'une séquence.

L'arbre phylogénétique inféré par la méthode de parcimonie sera donc celui rendant compte du minimum d'évènements évolutifs ayant pu se produire et donc du chemin le plus court permettant d'aller d'un taxon à un autre (Fitch 1971), soit le plus petit nombre de substitutions possible entre les données. Parmi toutes les topologies d'arbres comparées, celle qui sera retenue sera celle de l'arbre ayant obtenu le plus petit score parcimonieux.

Tout comme les méthodes de distances décrites ci-dessus, l'analyse en maximum de parcimonie est sujette au biais de l'attraction des longues branches. De plus, comme Felsenstein l'a montré (Felsenstein 1978), cette méthode peut parfois être inconsistante, c'est-à-dire amener à un résultat erroné (en statistiques, la notion de consistance prétend que la probabilité d'un paramètre doit tendre vers 1 lorsque le nombre de données tend vers l'infini).

Dans l'étude que nous avons réalisée sur l'évolution du virus de la Peste porcine africaine, nous avons souhaité éviter les biais d'analyses induits par les méthodes de distances et de maximum de parcimonie. Aussi n'avons-nous employé que les méthodes probabilistes de maximum de vraisemblance et d'inférence bayésienne avec technique de Monte Carlo, méthodes que nous allons donc davantage détailler dans les chapitres suivants.

2-6- Les méthodes probabilistes

Les méthodes phylogénétiques probabilistes reposent sur le concept de vraisemblance. Le concept de vraisemblance a été introduit par R.A. Fisher en 1922 (Aldrich 1997). La vraisemblance est une probabilité qui a pour objectif d'expliquer un jeu de données étudié selon un modèle probabiliste donné particulier. La valeur de cette probabilité peut alors être considérée comme l'expression de l'adéquation entre le modèle choisi et le jeu de données auquel on l'a appliqué. Les méthodes probabilistes utilisent des modèles évolutifs pour expliquer la phylogénèse d'organismes via les séquences d'ADN ou de protéines et indiquent donc si le modèle choisi est plus ou moins adapté pour résoudre le jeu de données étudié. Les modèles d'évolutions moléculaires, appelés aussi modèles de substitution, tentent de décrire le processus stochastique à l'œuvre lors de la survenue de

substitutions au sein de séquences tant d'ADN que de protéines, et ainsi d'expliquer les divergences observées entre des séquences homologues. Ils ont été développés au cours du temps et prennent en compte un nombre plus ou moins important de paramètres, paramètres déterminés d'après les observations des processus biologiques intervenant au cœur des cellules.

Ces modèles tiennent donc une place prépondérante dans la phylogénie moléculaire, puisqu'ils sont utilisés comme postulat initial par les méthodes probabilistes, tant de vraisemblance que d'inférence bayésienne. Il existe donc de nombreux modèles dont les conditions d'utilisation sont particulières et différentes en fonction des données étudiées. En effet, l'application d'un mauvais modèle, c'est-à-dire non adapté à un jeu de données entrainera une évaluation erronée des distances génétiques et donc produira un arbre phylogénétique éloigné de la réalité des liens unissant les taxons qui le compose.

Nous allons maintenant exposer les modèles évolutifs les plus courants que nous avons employés pour résoudre les relations entre isolats de virus de la PPA, un des objectifs de cette étude.

2-6-1- Les modèles évolutifs

L'analyse de la divergence, et inversement des similitudes, existant entre des séquences passe obligatoirement par la détermination de la distance qui les sépare. La divergence observée ou *p-distance* n'est pas suffisante en elle-même pour rendre compte du chemin évolutif qui a été franchi au cours du temps et qui explique cette divergence. En effet, dans le cas où une séquence d'ADN aurait une composition homogène, les quatre bases qui la composent auraient une fréquence de 0,25. Ainsi, la comparaison entre deux séquences donnerait, elle aussi, une proportion de résidus identiques égale à 0,25. Or, ceci n'est jamais biologiquement vérifié, notamment en raison du phénomène des substitutions multiples ou des réversions.

Pour compter le nombre X de mutations ayant eu lieu au cours du temps t, il faut donc considérer qu'à chaque instant t il a pu se produire un évènement de substitution. La probabilité qu'un tel phénomène se soit reproduit n fois au cours de t et à un taux μ est donc : $P_n(t) = P(X(t) = n)$. Bien évidemment, cette probabilité change si t change, et d'autant plus si le μt, qui est le nombre de substitutions par unité de temps, est élevé. Il est donc possible de calculer la probabilité qu'une mutation intervienne ou non, durant un intervalle de temps δt s'étant produit après t. Cette probabilité sera fonction de $\mu\delta t$.

Au final, la probabilité qu'une mutation intervienne au cours d'un temps t peut être formulée comme suit : $P_n(t) = [(\mu t)^n \exp(-\mu t)]/n!$, qui suit une distribution de Poisson, c'est-à-dire que le nombre de mutations n ayant eu lieu jusqu'au temps t suit une distribution de Poisson ayant pour paramètre μt.

2-6-2- Modélisation des substitutions selon un processus homogène markovien

Nous venons de le voir, le processus de substitution se réalisant au sein d'une séquence d'ADN suit une loi de Poisson, mais n'est cependant pas suffisant pour estimer correctement la distance évolutive réelle entre deux séquences. Il convient donc d'utiliser des modèles qui modulent le calcul de la *p-distance* en fonction des phénomènes biologiques à l'œuvre. Pour ce faire, l'on va utiliser le modèle des chaines de Markov. Dans les chaines de Markov, les substitutions à chaque site d'une séquence représentent les états possibles de la chaine et sont donc au nombre de quatre pour une molécule d'ADN, soit les quatre nucléotides de l'alphabet nucléique (ils seront au nombre de 20 pour les séquences protéiques). La probabilité P_{ij} du passage d'un nucléotide i vers un nucléotide j au cours du temps, avec $i, j \in \{A, T, C, G\}$ décrit l'évolution des sites entre deux séquences d'ADN, mais, le processus de Markov étant dit « sans mémoire », l'état d'un caractère dépend uniquement de son prédécesseur direct, sans rappel d'un passé plus lointain. L'état futur de ce même caractère ne dépendra donc que de son état présent. Les modèles de Markov, pour calculer les probabilités de passage d'un nucléotide i vers un nucléotide j utilisent une matrice (dite matrice Q) (Figure 5), spécifiant le taux relatif de changements de chaque nucléotide tout le long des séquences étudiées. L'utilisation d'une telle matrice classe ces modèles en tant que modèles markoviens homogènes et stationnaires, car ils répondent à plusieurs postulats : (i) à chaque site d'une séquence, le taux de remplacement d'une base par une autre est indépendant de la base qui occupait auparavant ce site, (ii) le taux de substitution μ_{ij} ne varie pas au cours du temps (homogénéité) et (iii) les fréquences relatives π_A, π_T, π_C et π_G des quatre nucléotides sont à l'équilibre dans les séquences (stationnarité).

A	C	G	T
$-\mu(a\pi_C + b\pi_G + c\pi_T)$	$a\mu\pi_C$	$b\mu\pi_G$	$c\mu\pi_T$
$g\mu\pi_A$	$-\mu(g\pi_A + d\pi_G + e\pi_T)$	$d\pi_G$	$e\pi_T$
$h\mu\pi_A$	$j\mu\pi_C$	$-\mu(h\pi_A + j\pi_C + f\pi_T)$	$f\mu\pi_T$
$f\mu\pi_T$	$k\mu\pi_C$	$l\mu\pi_G$	$-\mu(i\pi_A + k\pi_C + l\pi_G)$

Figure 5 : Matrice $\boldsymbol{Q}$ déterminant le taux de substitution instantané d'un nucléotide $\boldsymbol{i}$ vers un nucléotide $\boldsymbol{j}$ lors d'un processus de Markov utilisé pour le modèle de substitution le plus complet, le modèle GTR. $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}, \boldsymbol{l}$ sont les taux relatifs de substitution d'un nucléotide par un autre. $\boldsymbol{\pi_A}, \boldsymbol{\pi_T}, \boldsymbol{\pi_C}, \boldsymbol{\pi_G}$ sont les fréquences observées de chaque nucléotides dans les séquences. La somme de la diagonale est égale à 0.

Bien évidemment, ces présupposés ne rendent pas totalement compte de la réalité biologique des forces évolutives à l'origine de la divergence des séquences d'ADN. En revanche, ils modélisent relativement bien les processus stochastiques qui sont à l'œuvre.

En plus de l'homogénéité, de la stationnarité et de l'indépendance des sites, les processus markoviens sont dits réversibles. Cela signifie qu'au cours du temps, pour chaque

état i ou j d'un caractère, $\pi_i \mu_{ij} = \pi_j \mu_{ji}$. Autrement dit, à l'état d'équilibre (état hypothétique atteint par une séquence après un temps infini d'évolution), le nombre de substitutions du caractère i vers le caractère j est identique au nombre de substitutions du caractère j vers le caractère i. Ce processus de réversibilité a une incidence très importante dans l'élaboration des modèles évolutifs utilisés en phylogénie moléculaire : la matrice Q utilisée par le processus de Markov pour calculer les taux de passage d'un caractère i vers un caractère j est une matrice à 12 entrées. Or, selon l'hypothèse de réversibilité, les taux μ_{ij} et μ_{ji} sont symétriques donc interchangeables. Ainsi, sur les 12 paramètres initiaux nécessaires pour le calcul, seulement six paramètres d'échangeabilité peuvent être pris en compte (soit les quatre transversions et les deux transitions biologiquement possibles, nommées *a*, *b*, *c*, *d*, *e* et *f*). Le modèle doit également considérer les fréquences d'équilibre π_i de trois des quatre nucléotides (la quatrième étant forcément déductible des trois premières, puisque $\sum_i \pi_i = 1$). Ainsi, le modèle évolutif le plus complexe, le modèle GTR (pour « *General Time Reversible* ») ne tient-il compte « que » de neuf paramètres, soit les six paramètres d'échangeabilité et les trois fréquences d'équilibre (Barry & Hartigan 1987 ; Lanave *et al.* 1984 ; Rodriguez *et al.* 1990 ; Tavaré 1986 ; Yang 1994a).

Le nombre de substitutions ayant eu lieu dans une séquence au cours d'un temps t s'obtient dans un modèle markovien homogène par le produit du taux d'évolution global λ et du temps t. Or l'hypothèse de stationnarité implique que $\lambda = \sum_i \pi_i \lambda_i$ soit la somme du produit des fréquences et du taux d'évolution de chaque nucléotide. Dans le cas où une séquence ancestrale produit deux séquences filles, la distance évolutive entre les deux séquences filles doit tenir compte non plus d'un seul temps t mais de $2t$, et son expression devient alors : $d = 2 \sum_i \pi_i \lambda_i t$.

2-6-3- Les principaux modèles évolutifs markoviens en phylogénie moléculaire

Le modèle évolutif le plus complexe, développé entre autre par Yang en 1994, comporte neuf paramètres. Il est cependant l'aboutissement du développement de modèles antérieurs, qui au fil du temps, ont pris en compte de plus en plus de paramètres. Sur le modèle des poupées russes, le modèle GTR imbrique progressivement chacun des modèles qui l'ont précédé, du plus simple au plus complexe.

Le modèle le plus simple fut développé en 1969 par T.H. Jukes et C.R. Cantor (Jukes 1969). Devenu JC69 dans le langage courant, il postule une seule fréquence $\pi_i = 0{,}25$ pour chacun des nucléotides, un taux unique d'échangeabilité λ (ou taux de conservation global instantané), soit *a* = *b* = *c* = *d* = *e* = *f* = 1 et un seul taux de substitution instantanée α pour chacun des changements possibles. Ceci a une implication directe, puisque tous les taux de substitutions μ_{ij} étant égaux à α, alors il est possible de calculer la probabilité qu'un

nucléotide soit remplacé, ou bien reste identique au sein d'une séquence. Ces deux probabilités sont calculées comme suit : $P_{ii}(t) = \frac{1}{4} + \frac{3}{4}\exp(-\mu t)$ pour un nucléotide invariant et $P_{ij}(t) = \frac{1}{4} - \frac{1}{4}\exp(-\mu t)$ lorsque le nucléotide a changé. On peut alors déduire la distance entre les séquences à l'équilibre, c'est-à-dire pour un temps $t \rightarrow \infty$ pour lequel $\pi_i = 0{,}25$, soit $d = 6\alpha t$. De là, la relation entre d et $P(t)$ peut être déduite, soit

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3}p).$$

Ce modèle n'était cependant pas satisfaisant, puisque la fréquence des transitions est supérieure à celle des transversions dans les substitutions nucléotidiques qui interviennent au cours du temps dans les séquences d'ADN. Or, le modèle JC69 présuppose des taux d'échangeabilité identiques. Tenant compte de cette observation biologique, Kimura a proposé un modèle à deux paramètres libres, intégrant un taux de transition α différent du taux de transversion β (Kimura 1980). Ainsi, dans ce modèle, le taux de substitution instantanée, qui est égal au taux de substitution global est $\lambda = \alpha + 2\beta$. Appelant $r(t)$ la probabilité d'observer une transition, $v(t)$ celle d'observer une transversion et $q(t)$ aucun changement de nucléotide après un temps t, le modèle K80 déduit que la relation entre la distance entre deux séquences et les probabilités d'apparition de substitutions (transitions ou transversions) à l'équilibre est la suivante : $d = 2(\alpha + 2\beta)t$ d'où on peut déduire la relation avec $r(t)$ et $v(t)$:

$$d = -\frac{1}{2}\ln(1 - 2r - v) - \frac{1}{4}\ln(1 - 2\,v)$$

Les deux modèles précédemment présentés, JC69 et K80, se fondent sur l'hypothèse qu'à l'état d'équilibre, les fréquences π_i des quatre nucléotides sont toutes égales à 1/4, ce qui revient à dire que le ratio GC des nucléotides dans une séquence soit égal à 1/2. Or, cette hypothèse n'est pas vérifiée biologiquement. Depuis Chargaff et la détermination paritaire A – T et G – C dans la molécule d'ADN, le GC% est devenu une mesure fréquente de l'hétérogénéité des séquences. Les liaisons G – C étant plus fortes que les liaisons A – T, la composition d'une séquence en G et en C a donc une importance évidente en termes de stabilité de la molécule. La détermination du GC% a montré que sa valeur variait de 25% à 75%, et que cette proportion pouvait également varier au sein d'un même organisme, selon les zones de l'ADN génomique considérées (Bernardi 1993a, b). Ces différentes proportions de GC dans un génome pourraient être seulement dues aux variations mutationnelles stochastiques produites lors de la réplication de l'ADN, mais d'aucun pense que ces proportions doivent conférer un avantage sélectif aux organismes qui les portent. Par exemple, il a été déterminé que les génomes riches en AT sont souvent de petite taille et appartiennent le plus souvent à des organismes parasitaires (Moran 2002 ; Rocha & Danchin 2002 ; Wang *et al.* 2006). A titre d'exemple, le virus de la PPA est riche en AT. La faiblesse des liaisons d'appariement entre les deux brins d'ADN relative à cette composition en nucléotides pourrait conférer au virus un avantage réplicatif, les hélicases, responsables de

la séparation des brins d'ADN, pouvant alors avoir une efficacité augmentée. Il a même été déterminé que des populations colonisant certaines niches écologiques voyaient la composition relative en nucléotides de leur génome affectée de façon spécifique par ces niches (Foerstner *et al.* 2005). Ceci est bien entendu en rapport avec les acides aminés des protéines codées par ces génomes, dont les interactions avec l'environnement sont, comme nous l'avons déjà vu, importantes pour la survie de ces organismes.

Pour pallier ce biais, des modèles évolutifs tenant compte de la diversité compositionnelle des séquences d'ADN ont été développés. En 1992, K. Tamura, à partir du modèle K80, a proposé un modèle à trois paramètres libres (T92), intégrant ces variations nucléotidiques (Tamura 1992). En plus des paramètres α et β précisant les taux de transitions et de transversions, il a ajouté le paramètre θ représentant la proportion en bases GC des séquences étudiées. Bien évidemment, dans le cas d'une séquence pour laquelle le GC% = 50, la valeur de θ sera égale à 0,5, et le modèle T92 sera alors pleinement équivalent au modèle K80. Dans tous les autres cas, c'est-à-dire la très grande majorité, le taux λ d'évolution des nucléotides sera défini comme suit :

$$\lambda_A = \lambda_T = -(\alpha\theta + \beta) \text{ et } \lambda_G = \lambda_C = (\theta - 1)\alpha - \beta$$

Les valeurs $r(t)$ de probabilité de transitions, $v(t)$ de transversions et $q(t)$ de non substitutions peuvent être déterminées, et les fréquences attendues à l'équilibre après un temps d'évolution infini ne sont plus égales à $\pi_A = \pi_T = \pi_C = \pi_G = 0{,}25$ mais $\pi_A = \pi_T = (1-\theta)/2$ et $\pi_C = \pi_G = \theta/2$. La distance évolutive entre deux séquences devient alors $d = 4\theta(1-\theta)\alpha t + 2\beta t$. De même que pour le modèle K80, il est possible de déduire de cette formule les liens unissant d, $r(t)$ et $v(t)$:

$$d = -h\ln\left(1 - \frac{r}{h} - v\right) - \frac{1}{2}\ln(1-h)\ln(1-2v) \text{ où } h = 2\theta(1-\theta)$$

Il est à noter que θ n'est bien évidemment pas fixe, et variera donc entre les séquences comparées.

Cependant, les modèles précédents ne tiennent pas compte de la nature des transitions entre les bases puriques ou les bases pyrimidiques. Pour tenir compte de cette réalité biologique, un modèle à six paramètres libres a été proposé en 1993 (Tamura & Nei 1993) : le TN93. Dans ce modèle évolutif, α_R et α_Y caractérisent respectivement les taux de transitions puriques (A↔G) et pyrimidiques (T↔C), tandis que β représente toujours le taux des transversions. Ce modèle fait apparaitre explicitement les fréquences π_i des nucléotides à l'équilibre, avec $\pi_R = \pi_A + \pi_G$ et $\pi_Y = \pi_T + \pi_C$. La valeur des taux de transitions $r(t)$, est donc scindée en deux, soit $r_R(t)$ et $r_Y(t)$ tandis que la valeur $v(t)$ de transversions reste inchangée. La distance entre deux séquences peut alors s'exprimer comme suit :

$$d = 2(\pi_A\pi_G\alpha_R + \pi_T\pi_C\alpha_Y + \pi_R\pi_Y\beta)2t = 4\pi_A\pi_G\alpha_R t + 4\pi_T\pi_C\alpha_Y t + 4\pi_R\pi_Y\beta t$$

La relation entre la distance (d) , $r_R(t)$, $r_Y(t)$ et $v(t)$ devient donc :

$$d = \frac{2\pi_T\pi_C}{\pi_Y}(\alpha_1 - \pi_R b) + \frac{2\pi_A\pi_G}{\pi_R}(\alpha_2 - \pi_Y b)$$

avec :

$$\alpha_1 = -\ln(1 - \frac{\pi_Y}{2\pi_T\pi_C}r_Y - \frac{1}{2\pi_Y}v)$$

$$\alpha_2 = -\ln(1 - \frac{\pi_R}{2\pi_A\pi_G}r_R - \frac{1}{2\pi_R}v)$$

$$b = -\ln(1 - \frac{1}{2\pi_R\pi_Y}v)$$

Les mécanismes sous-jacents à l'évolution des séquences d'ADN, même s'ils sont stochastiques, peuvent parfois amener à des cas particuliers, comme par exemple des séquences dans lesquelles α_R et α_Y soient égaux ou encore $\alpha_R = \beta(1 + \kappa/\pi_R)$ et $\alpha_Y = \beta(1 + \kappa/\pi_Y)$ où κ est le rapport entre le taux de transitions et le taux de transversions. Pour modéliser ces cas particulier, des dérivés du modèle TN93 ont été proposés. Le modèle HKY85 a été développé (Hasegawa *et al.* 1985) pour répondre à la parité α_R et α_Y tandis que le modèle F84 a été proposé pour modéliser le second cas particulier (Felsenstein & Churchill 1996). Ces modèles ont donc un paramètre de libre de moins que leur modèle précédent, c'est-à-dire cinq paramètres libres.

L'imbrication et la hiérarchisation de ces modèles évolutifs ainsi que les paramètres qu'ils prennent en compte est résumé dans la figure 6 ci-dessous.

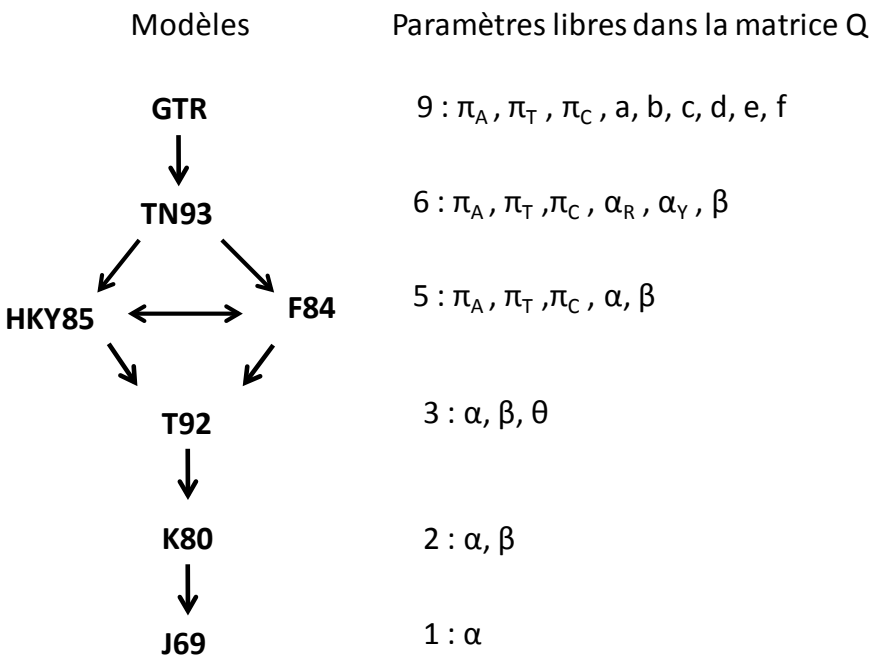


Figure 6 : Imbrication des différents modèles évolutifs markoviens. Le nombre et la nature des paramètres libres dans la matrice markovienne $\boldsymbol{Q}$ sont indiqués. Dans le modèle JC69, $\boldsymbol{\alpha}$ est le taux de substitution observé entre deux séquences. Dans le modèle K80, $\boldsymbol{\alpha}$ est le nombre de transitions et $\boldsymbol{\beta}$ le nombre de transversions. Pour le modèle T92, $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ sont les mêmes que pour le modèle K80, et $\boldsymbol{\theta}$ le pourcentage de bases GC dans les séquences. Pour les modèles HKY85 et F84, $\boldsymbol{\pi_A}$, $\boldsymbol{\pi_T}$, et $\boldsymbol{\pi_C}$ sont les fréquences observées des bases A, T et C, $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ restant les mêmes. Dans le modèle TN93, outre les fréquences des bases A, T, C et les transversions $\boldsymbol{\beta}$, les transitions sont séparées en $\boldsymbol{\alpha_R}$ pour les transitions entre deux purines et $\boldsymbol{\alpha_Y}$ les transitions entre bases pyrimidiques. Enfin, le modèle GTR tient compte des fréquences A, T et C ainsi que des six possibilités de substitutions possibles (2 transitions et 4 transversions) : **a**, **b**, **c**, **d**, **e** et **f**.

2-7- Modèles de codons pour les séquences codantes

Les modèles évolutifs assument pour la plupart que les nucléotides évoluent indépendamment les uns des autres. Or, cette assertion n'est pas réellement validée d'un point de vue biologique, principalement en ce qui concerne les séquences codantes. Dans le cas des séquences codantes, en effet, la position du nucléotide à l'intérieur du codon n'est pas sans importance. De par la dégénérescence du code génétique, une substitution en position 1, 2 ou 3 aura un impact différent en termes de protéine traduite. En effet, dans des séquences codantes, les substitutions synonymes sont assez nombreuses, et informent sur la diversité récente, tandis que les substitutions non synonymes, plus rares, fournissent des informations sur la diversité précoce entre les séquences (Ren *et al.* 2005). On observera donc classiquement dans une séquence un taux élevé de substitution en position 3 du codon, alors que ce taux sera réduit sur la position 2 (Yang 1996). Aussi, des modèles prenant en compte le codon dans son entier ont été développés. Dans ces modèles l'unité de substitution n'est plus le nucléotide, à une position donnée, mais le codon, dans lequel un seul nucléotide peut être substitué, par unité de temps (Goldman & Yang 1994 ; Muse &

Gaut 1994 ; Pedersen *et al.* 1998). Pris en tant qu'unité évolutive, il est possible de distinguer les substitution synonymes des substitutions non synonymes (Ren *et al.* 2005). Ainsi, le paramètre $\omega = dN/dS$, qui représente le taux de substitutions synonymes sur le taux de substitutions non synonymes par unité de temps, pourra être calculé lors de l'analyse, et pris en compte lors de l'inférence de la phylogénie des séquences étudiées. En intégrant à la fois la divergence nucléotidique et le code génétique, les modèles peuvent alors incorporer aux calculs des paramètres jusqu'ici ignorés, à savoir la non indépendance des sites voisins et les différences de taux d'évolution entre sites d'un même codon (Goldman & Yang 1994). La prise en compte de ces informations permet de mieux établir les liens phylogénétiques entre les organismes étudiés.

Les modèles courants utilisant les codons comme unité d'évolution sont basés sur un matrice à 61 entrées, correspondant aux 61 codons sens du code génétique, qui permet de calculer les probabilités de substitution entre deux codons. Cette matrice peut être générée selon deux principales approches : les modèles paramétriques dérivant la matrice de l'observation directe des séquences, en intégrant, par exemple, le ratio taux de transitions/taux de transversions (κ), ainsi que ω (Goldman & Yang 1994), ou en explicitant les fréquences des nucléotides (Muse & Gaut 1994). Dans des modèles ultérieurs, ω pourra être amené à varier entre les lignées (Yang 1998), ou entre les sites (Nielsen & Yang 1998). Les méthodes empiriques, plus récentes, basent l'estimation de la matrice sur l'observation de grands jeux de données, en autorisant par exemple, les substitutions multiples instantanées (Schneider *et al.* 2005), qui correspondent aux nombres de substitutions nécessaires pour passer d'un codon à un autre. Des modèles semi-empiriques ont depuis été développés, dans lesquels sont utilisés à la fois une matrice empirique et des paramètres de l'analyse paramétrique (Doron-Faigenboim & Pupko 2007 ; Kosiol *et al.* 2007). Les paramètres les plus pertinents dans une analyse « codons » ont été identifiés comme étant : ω, κ, et le taux de substitutions multiples υ (Zoller & Schneider 2010). Ces modèles semi-empiriques ont été intégrés dans les analyses en datation moléculaire, comme c'est le cas avec le logiciel BEAST (Drummond & Rambaut 2007), que nous avons utilisé pour notre étude.

2-8- Maximum de vraisemblance

Le concept de vraisemblance tend à aider à la prise d'une décision quant à la valeur prédictive de l'application d'un modèle probabiliste pour expliquer un jeu de données. Comme son nom l'indique, il tente de déterminer si le modèle utilisé est le plus vraisemblable, c'est-à-dire le plus plausible. Ce genre de prise de décision est typiquement ce que l'on est amené à faire lorsqu'on dispose de données biologiques à analyser. Dans le contexte de l'analyse phylogénétique, la question de plausibilité repose sur la façon dont un arbre représente réellement les relations entre des taxa. Le questionnement concerne la

topologie de l'arbre, la longueur des branches qui le compose, ou encore le modèle évolutif choisi pour le construire. Autant d'hypothèses qui, si elles s'avèrent erronées, conduiront à de fausses conclusions. On peut donc dire que, dans le contexte phylogénétique, la valeur de vraisemblance attribuée à un arbre atteste de la fiabilité du décryptage du signal phylogénétique contenu dans le jeu de données étudié (Huelsenbeck 2001).

La vraisemblance propose donc de calculer la probabilité conditionnelle L d'observer des données de séquences S selon une hypothèse H. Cette probabilité conditionnelle s'exprime comme suit : $L_S = Pr(S|H)$. La vraisemblance est donc une fonction de modèles paramétrés qui propose une fonction de densité $f(x\,;p)$ dans laquelle p est un vecteur de paramètres. Typiquement, en phylogénie moléculaire, p représentera la longueur des banches de l'arbre ainsi que le taux μ de substitution par site le long d'une séquence, et x les échantillons indépendants testés. En d'autres termes, la vraisemblance va donc statuer sur la probabilité $L \in [0,1]$ de plausibilité de la résolution du jeu de données selon un modèle. Elle va maximiser H pour chaque scenario envisagé, et celui ayant obtenu la probabilité L la plus élevée sera le scenario retenu, selon le principe de *maximum de vraisemblance*.

La méthode de maximum de vraisemblance appliquée à la phylogénie moléculaire a été initialement suggérée en 1967 (Cavalli-Sforza & Edwards 1967), mais la lourdeur des calculs nécessaires à son utilisation en limitait pratiquement l'usage. Dès 1971, le principe de maximum de vraisemblance fut pourtant appliqué aux séquences nucléotidiques et protéiques en postulant que les changements observés se faisaient aléatoirement et indépendamment les uns des autres (Neyman 1971). Par la suite, c'est essentiellement Felsenstein qui développera cette méthode, d'abord en 1973 (Felsenstein 1973a, b), puis surtout en 1981 (Felsenstein 1981) grâce au développement de l'algorithme d'élagage (ou « pruning ») qui permettra de calculer la vraisemblance pour un grand nombre de séquences. En effet, le nombre de nœuds internes d'un arbre pour un site considéré de la séquence d'un taxon croît exponentiellement avec le nombre de taxa étudiés : pour n taxa étudiés, il y a 4^{n-1} possibilités, soit autant de calcul de probabilité à effectuer. Par exemple, pour analyser un jeu de données de 50 taxons, le nombre de termes est donc de 4^{49}, soit environ 3 x 10^{29}. L'algorithme d'élagage retire deux feuilles de l'arbre à chaque étape, permettant ainsi de limiter grandement le nombre de calculs nécessaire pour construire l'arbre. On notera que l'élagage produit une vraisemblance partielle pour chaque nœud K des sous-arbres générés. Elle est notée $L_K^i(k)$, où k est l'état observé du nœud pour le site considéré. De fait, cette probabilité correspond donc à la vraisemblance du sous-arbre raciné en K, sachant k.

Les séquences homologues étudiées en phylogénie moléculaire sont le résultat d'une histoire évolutive inconnue. Néanmoins, cette histoire évolutive tient compte de divers paramètres, ou vecteur de paramètres (Θ). Ces paramètres comprennent les relations de filiation des séquences représentées par la topologie de l'arbre (τ), la quantité d'évolution entre les séquences ayant permis le calcul de la longueur des branches de l'arbre (b), ainsi

que l'ensemble des paramètres définis par le modèle évolutif markovien appliqué tentant de déterminer les processus évolutifs à l'œuvre pour expliquer les données, soit (θ). Il est évident que le nombre tant des arbres que des mécanismes de substitutions ayant conduit à l'état des caractères observés aujourd'hui est excessivement élevé. Toutefois, parmi tous les scenarii évolutifs possibles, certains sont plus vraisemblables que les autres, c'est-à-dire qu'ils retracent mieux la réalité évolutive des séquences étudiées. Le calcul du maximum de vraisemblance va donc avoir pour but de déterminer le Θ optimal permettant d'expliquer les observations. Cette probabilité s'écrit $P(S|\Theta)$ (signifiant la probabilité de S sachant Θ), et revient au final à (i) tester et estimer un ensemble d'hypothèses, et (ii) à réaliser une reconstruction phylogénétique permettant de déterminer le Θ optimal, c'est-à-dire de produire l'arbre dont la topologie τ sera la plus vraie. Le modèle évolutif sous-jacent employé pour réaliser ces calculs statistiques devra donc être celui qui correspond le mieux au jeu de données de séquences étudié pour que les biais de calculs soient minimisés.

L'objectif des calculs de vraisemblance est donc de déterminer la probabilité qu'une séquence donnée ait donné naissance à une ou plusieurs autres séquences au cours du temps, en calculant la probabilité de l'ensemble des paramètres de l'histoire évolutive des séquences comparées. Pour simplifier ces calculs, l'hypothèse de l'indépendance des données est posée. Ainsi, chaque site i d'une séquence est considéré comme indépendant des autres et il en est de même entre les différentes séquences. Tenant compte de ces deux hypothèses, la vraisemblance globale $L(\Theta)$ d'une séquence sera le produit des vraisemblances $L^{(i)}(\Theta)$ calculée à chaque site i de la séquence. Elle s'exprime donc selon la formule suivante :

$$L(\Theta) = P(S|\Theta) = \prod_{i=1}^{l} L^{(i)}(\Theta) = \prod_{i=1}^{l} P\left(S^{(i)}\middle|\Theta\right)$$

où $S^{(i)}$ est un site quelconque i de l'alignement (soit une colonne) et l la longueur totale des séquences, c'est-à-dire le nombre de sites considérés.

L'algorithme d'élagage développé par Felsenstein génère des vraisemblances partielles au niveau des différents nœuds, comme au niveau des feuilles de l'arbre. Au niveau des feuilles, la vraisemblance est factuelle, puisque se basant sur l'observation du site tel qu'il se présente dans le jeu de données disponibles basé sur des séquences réelles observées. Ainsi, la vraisemblance d'une feuille au nœud K (ou nœud externe) sera $L_K^i(k) = 1$ pour le nucléotide observé au site i, et $L_K^i(k) = 0$ pour les trois autres nucléotides, un site ne pouvant en effet n'être constitué que par un seul état du caractère « nucléotide ». En revanche, pour calculer la vraisemblance partielle d'un nœud interne K de l'arbre, c'est-à-dire l'hypothèse conditionnelle d'une séquence ancestrale des séquences réelles observées, il faut au préalable avoir calculé la vraisemblance partielle de ses deux nœuds fils (rappelons ici que la phylogénie probabiliste fonctionne sur un mode de bifurcation, les modèles autorisant les polytomies, c'est-à-dire les réseaux, seront développés ultérieurement). Tous

les calculs ont donc pour origine les feuilles des arbres, soit les séquences réelles observées, représentant les nœuds ultimes de l'arbre.

Si un nœud K a deux nœuds fils L et M, il faudra donc considérer tous les états possibles des caractères m et l menant au caractère k du nœud K. La vraisemblance de K sera donc égale au produit des vraisemblances des états m et l. Or, dans une séquence d'ADN, quatre états sont possibles : A, T, C ou G. Le calcul de vraisemblance pour chacun des états possibles sera donc effectué, calcul qui fournira la probabilité d'avoir l'état l au nœud L, par exemple, pour le site considéré. Cette probabilité sera calculée puis sommée pour chacun des caractères de la séquence L. Il en sera de même pour la séquence M et ainsi, la probabilité que L et M mène à K sera le produit de ces sommes. Elle est alors formulée comme suit :

$$L_K^i(k) = \left[\sum_l P(l|k, b_l) L_L^{(i)}(l)\right] . \left[\sum_m P(m|k, b_m) L_M^{(i)}(m)\right]$$

avec b la longueur de la branche menant au nœud considéré.

Les probabilités partielles seront ainsi calculées, pour chaque site des séquences étudiées, en partant des feuilles pour aller jusqu'à la racine de l'arbre. La probabilité d'avoir un état de caractère I à chaque site i de la séquence racine correspondant donc à la somme des probabilités de chaque site à chaque nœud fils de l'arbre sur la longueur l des séquences.

L'arbre ainsi généré sera donc le plus vrai possible, « probabilistiquement » parlant. Il reste tout de même un écueil à cette méthode, c'est qu'elle présuppose un taux d'évolution constant entre toutes les branches, c'est-à-dire pour chaque site et tout le long de la séquence, ce qui est contestable sur le plan biologique. C'est pourquoi Yang a proposé un modèle tenant compte d'un taux variable d'évolution selon les sites (Yang 1994b).

Si la méthode probabiliste par maximum de vraisemblance s'est généralisée chez les phylogénéticiens, c'est qu'elle est non biaisée (l'estimation de la probabilité du paramètre étudié est vraie dans les conditions asymptotiques, c'est-à-dire pour les grandes quantités de données), efficiente (l'estimation de Θ est optimale au maximum de vraisemblance) et consistante (la valeur de Θ augmente avec le nombre de données analysées).

Outre la fiabilité de la reconstruction phylogénétique, la méthode de maximum de vraisemblance permet également dans un second temps, de tester statistiquement ses résultats. Comme l'a montré Felsenstein, les estimations de Θ suivent une distribution normale autour de la vraie valeur de Θ (Felsenstein 2004). Cela permet de calculer des intervalles de confiances des estimations faites, car les différences entre les valeurs de vraisemblance obtenues pour plusieurs hypothèses testées doivent suivre une loi du χ^2, le nombre de degrés de liberté étant fonction du nombre N de données. Cependant, si les valeurs de Θ ne suivent pas parfaitement une distribution normale autour de la vraie valeur

de Θ, l'intervalle de confiance calculé sera faux, pouvant donc éventuellement confirmer une hypothèse fausse. Pour pallier ce problème, le test des ratios de vraisemblance LRT (pour *likelihood ratio test*) a été développé (Goldman 1993). En effet, le double de la différence des LRT entre deux hypothèses a été démontré comme suivant lui aussi une loi du χ^2 à k degrés de liberté (k correspondant à la différence du nombre de paramètres entre les deux modèles testés (Posada 1998 ; Yang 2006), soit $2[\ln L(\Theta_1) - \ln(\Theta_0)] \sim \chi^2_k$. Ce test interviendra, nous le verrons, dans le test de l'hypothèse de l'horloge moléculaire stricte implémenté dans le logiciel PAML lors des analyses de datation moléculaire que nous avons réalisées (cf. infra).

De même, des critères d'informations ont été développés pour aider au choix du modèle le plus en adéquation avec le jeu de données étudiées, tel l'Akaike Information Criterion (AIC), l'Akaike Information Criterion critical (AICc) ou encore le Bayesian Information Criterion (BIC). Nous ne les développerons pas ici puisqu'ils sont expliqués dans le chapitre Matériels et Méthodes du présent manuscrit.

2-9- Inférence bayésienne – chaines de Markov et technique de Monte Carlo

L'expression de la vraisemblance d'une hypothèse s'exprime par la formule $P(S|\Theta)$, soit la probabilité de S sachant Θ définissant les paramètres impliqués dans le modèle M. Cette probabilité est donc l'expression *a posteriori* de l'application d'un modèle évolutif à un jeu de données. Or, la valeur de la vraisemblance d'une hypothèse n'atteste pas la véracité de cette hypothèse. Pour vérifier cela, il est envisageable d'inverser les termes de la probabilité, de façon à déterminer Θ (les paramètres du modèle M), sachant S, ce qui s'écrit $P(\Theta|S)$. En d'autres termes, cela revient à pondérer *a priori* les hypothèses évolutives testées.

Les probabilités conditionnelles, incluant le théorème que Bayes a développé au XVIIIème siècle, n'ont été intégrées que récemment dans les analyses de phylogénie moléculaire, car les calculs de probabilités postérieures reposaient sur des intégrales multidimensionnelles que seuls les dernières générations de calculateurs pouvaient résoudre. Elles font désormais partie de l'arsenal classique des analyses phylogénétiques. Le théorème que Bayes développa est le suivant :

$$P(\Theta|S,M) = \frac{P(S|\Theta,M)P(\Theta|M)}{P(S|M)}$$

C'est-à-dire qu'il permet la détermination de la probabilité $P(\Theta|S,M)$, qui est alors nommée probabilité *a posteriori*, exprimée comme la probabilité de réalisation de Θ du modèle M, sachant S. A la différence du calcul « simple » de la vraisemblance, dans lequel toutes les hypothèses sont considérées *a priori* comme équiprobables (Kuhner *et al.* 1995),

le théorème de Bayes va permettre d'affecter à chacune de ces hypothèses une probabilité *a posteriori* afin de valider ou d'infirmer leur plausibilité. En résumé, le théorème de Bayes appliqué à la phylogénie va permettre d'ajouter à la valeur de vraisemblance d'un jeu de données sachant le modèle utilisé pour la résoudre (probabilité *a posteriori*), la valeur de vraisemblance *a priori* de la résolution de ce jeu de données par le modèle considéré. Le calcul des intégrales multidimensionnelles pour évaluer la distribution des espérances *a posteriori* des paramètres était impossible jusqu'à l'implémentation de méthodes numériques, ce qui empêchait l'utilisation des probabilités conditionnelles en phylogénie moléculaire. La plus usitée des méthodes numériques est celle des « chaines de Markov avec technique de Monte Carlo », ou MCMC pour *Monte Carlo Markov Chain*. Le concept des chaines markoviennes a été introduit précédemment lors de la description des modèles évolutifs (cf. supra), en expliquant qu'une chaine de Markov permettait un échantillonnage prenant la forme d'une distribution stationnaire (c'est-à-dire à l'équilibre), en se projetant dans l'espace multidimensionnel des paramètres impliqués dans les modèles évolutifs appliqués à la résolution des données de séquences. La technique de Monte Carlo implémentée dans une chaine de Markov va permettre d'approximer la distribution *a posteriori* pour venir compléter le modèle bayésien.

Les techniques de Monte Carlo sont basées sur le concept de simulation, utilisant des tirages aléatoires permettant le calcul de quantités déterministes. Elles répondent à la théorie des jeux suivant la loi des grands nombres, d'où son nom de Monte Carlo, en rapport avec les casinos qui peuplent ce rocher (Metropolis & Ulam 1949). En effet, pour étudier voire expérimenter un système basé sur des interactions complexes, ce qui est particulièrement le cas de l'évolution des séquences d'ADN, la simulation permet de mesurer les changements induits par la modification de certains paramètres sur le comportement du système dans son entier et d'expérimenter des situations inconnues, aléatoires, voire improbables. Lorsque l'on désire modéliser fidèlement des évènements observables, l'on est assez rapidement confronté à des calculs non explicites, souvent inaccessibles, insolubles, parfois même d'un point de vue conceptuel. Les techniques de Monte Carlo, grâce aux simulations, vont ainsi permettre une approximation de ces calculs. Pour ce faire, ces techniques vont faire appel à la répétition des expériences, dont les résultats, cumulés et comparés les uns aux autres, vont pouvoir évaluer la fiabilité de l'hypothèse première, grâce à la quantité de résultats générés, et ainsi pouvoir résoudre des systèmes pourtant déterministes et donc non réellement stochastiques.

Le nombre d'arbres générés lors de l'analyse en maximum de vraisemblance croît, nous l'avons dit, de façon exponentielle avec les nombre de séquences traitées. De plus, il ne suffit pas de compiler l'ensemble des arbres pour trouver la meilleure topologie, mais il faut également sommer les longueurs des branches de l'arbre, ainsi que les paramètres du modèle évolutif qui a été utilisé pour résoudre le jeu de données. Les chaines markoviennes, puisqu'elles comparent statistiquement leur état après chacun de leur tour, permettent de rejeter des hypothèses au fur et à mesure, pour tendre vers l'état dont la probabilité

postérieure sera la plus élevée. Cependant, la méthode de rejet, si la probabilité de rejet est trop grande, peut conduire à des simulations trop lentes, car elles demanderont un nombre de calculs, c'est-à-dire de tours de la chaine, bien trop important. L'application des techniques de Monte Carlo aux chaines markoviennes va justement pallier ce problème.

Rappelons d'une chaine de Markov qu'elle est une suite de variables aléatoires $\{X_n, n \geq 0\}$ dont l'état X_{n+1} ne dépend des valeurs passées que par X_n, c'est-à-dire son prédécesseur immédiat. La suite peut alors s'exprimer $X_{n+1} = f(X_n, Y_n)$, pour laquelle Y_n formerait une suite indépendante. La technique de Monte Carlo va ainsi permettre un échantillonnage aléatoire des éléments de cette suite, c'est-à-dire des espérances des probabilités postérieures des paramètres afin d'en déterminer la distribution, devenue alors discrète et non plus continue. C'est cette discrétisation qui permet de résoudre l'intégrale multidimensionnelle dont le calcul doit normalement donner la distribution des espérances. Si l'échantillonnage est suffisamment grand, c'est-à-dire si le nombre de tirages aléatoires est important, la théorie des grands nombres autorise à dire que le résultat convergera vers la vérité analytique. Cela revient à dire que la distribution de l'espérance mathématique d'une fonction g appliquée à une variable X, $g(x)$, et dont la fonction de densité est fx est désormais calculable grâce à la discrétisation due à l'échantillonnage de cette fonction de densité.

Appliquées à la phylogénie moléculaire, la variable aléatoire X correspond au vecteur de paramètres Θ tel que défini précédemment, c'est-à-dire représentant l'ensemble des paramètres définissant un modèle évolutif. Ainsi, la distribution de la variable aléatoire que l'on cherche à définir correspond-elle à la distribution postérieure des paramètres du modèle considéré et donc du modèle lui-même. Aussi, si les chaines MCMC sont suffisamment longues, l'approximation de l'espérance de la fonction $g(\mathrm{x})$, soit $E(g(\Theta)|S, M)$ sera-t-elle très proche de l'espérance « vraie ». L'arbre phylogénétique inféré pouvant donc être considéré lui aussi, comme le plus vrai selon les données et le modèle évolutif appliqué.

Terminons notre exposé de la méthode MCMC par l'approche bayésienne par échantillonnage Monte Carlo de chaines de Markov couplé à l'algorithme de Metropolis – Hasting, dit MCMCMC (pour *Metropolis Coupling Markov Chain Monte Carlo*) (Hastings 1970 ; Metropolis 1953). C'est cet algorithme qui est utilisé en datation moléculaire par le logiciel BEAST. Les MCMC permettent de générer un échantillonnage de leurs réalisations, distribué selon la loi de probabilité *a posteriori* du modèle évolutif. Cela permet donc, à partir d'un vecteur de paramètres Θ de produire le vecteur Θ', version modifiée de Θ. L'algorithme de Metropolis – Hasting permet, quant à lui, de comparer Θ et Θ' afin d'accepter ou de refuser ce nouvel état Θ', en fonction de sa probabilité conditionnelle *a posteriori*. C'est pourquoi, il est dit algorithme de rejet. La probabilité d'acceptation d'un nouveau vecteur Θ' peut être exprimée comme suit :

$$P_{accepter}(\Theta'|\Theta) = min\left(1, \frac{P(\Theta'|S,M)}{P(\Theta|S,M)} \frac{q(\Theta', d\Theta)}{q(\Theta, \mathrm{d}\Theta')}\right)$$

formule de probabilité conditionnelle dans laquelle $(\Theta, d\Theta')$ indique le mouvement du vecteur Θ vers le vecteur Θ' et $(\Theta', d\Theta)$ le mouvement inverse. Notons que $(\Theta, d\Theta')$ est appelé noyau stochastique de la chaine MCMC. Cette probabilité d'acceptation ou de rejet d'une hypothèse porte le nom de Metropolis – Hasting car le terme $\frac{P(\Theta'|S,M)}{P(\Theta|S,M)}$ correspond au ratio des probabilités *a posteriori* du mouvement $(\Theta, d\Theta')$ développé par Metropolis en 1953 tandis que le terme $\frac{q(\Theta', d\Theta)}{q(\Theta, \mathrm{d}\Theta')}$ représente lui, la pobabilité de l'aller-retour $\Theta \leftrightarrow \Theta'$, c'est-à-dire la probabilité $\Theta' \rightarrow \Theta$ (retour) sur la probabilité $\Theta \rightarrow \Theta'$ (aller) et a été développée par Hasting en 1970. Ainsi, si la modification d'un des paramètres du vecteur Θ a permis l'amélioration de la probabilité postérieure calculée pour le modèle considéré, le produit des ratios Metropolois – Hasting $\frac{P(\Theta'|S,M)}{P(\Theta|S,M)} \frac{q(\Theta', d\Theta)}{q(\Theta, \mathrm{d}\Theta')}$ sera supérieur à 1. Le nouvel état Θ'sera alors accepté avec une probabilité égale à 1, tandis que si cette modification du paramètre a entrainé une déterioration de la probabilité postérieure affectée au modèle, la probabilité de Θ' sera égale au produit des ratios Metropolis – Hasting et donc rejetée. La chaine MCMC continuera ainsi son chemin de proche en proche pendant un nombre de génération déterminé, garantissant que la probabilité $P(\Theta|S,M)$ sera toujours la meilleure et qu'ainsi l'arbre phylogénétique produit sera le meilleur sous le modèle considéré. Idéalement, la convergence des arbres ainsi générés au cours de l'analyse doit produire des valeurs de probabilité entre l'état Θ et l'état Θ' qui tendent vers zéro, les arbres étant de plus en plus semblables. Cependant, la complexité des liens unissant certains isolats peut amener le modèle à ne pouvoir trancher entre deux patrons relationnels. Les valeurs de probablité calculées après chaque tour des MCMC ne pourront alors pas devenir nulles mais oscilleront autour d'une valeur supérieure à zéro, ceci indiquant que l'optimum des calculs a été atteint.

3- La datation moléculaire

L'histoire de la datation moléculaire débute dans les années 1960, avec la parution de deux articles de E. Zuckerkandl et L. Pauling (Zuckerkandl & Pauling 1962, 1965). Ces deux auteurs, qui étudient l'évolution entre protéines homologues mais provenant d'espèces ayant divergé depuis longtemps, constatent une augmentation linéaire de la distance génétique en fonction du temps de spéciation. Ils concluent, peut-être un peu rapidement, à l'existence d'un taux d'évolution unique et constant au cours du temps, pour chaque classe de protéine (et des gènes qui les sous-tendent). L'hypothèse de l'horloge moléculaire était en gestation. Si elle s'avérait vraie, cette assertion avait une conséquence immédiate. Il devenait possible de dater avec précision la spéciation de chacune des espèces appartenant au règne du vivant, en calculant le taux d'évolution.

C'est Motoo Kimura, qui donne définitivement naissance à l'hypothèse de l'horloge moléculaire, en même temps qu'à la théorie de la neutralité évolutive (Kimura 1968). Il l'érigera en principe, en postulant que, tant que la structure tertiaire et la fonction des

protéines ne sont pas altérées, le taux de substitution des acides aminés reste constant, par site, et par unité de temps. Il modélisa cette théorie en partant du postulat que la fréquence d'émergence de tout nouvel allèle ne contrevenant pas au sélectionnisme darwinien, devait être uniquement due à une mutation accidentelle apparue entre deux générations. Ainsi, Kimura montra, mathématiquement, qu'un taux de substitution neutre pour des allèles, c'est-à-dire sans conséquence physiologique, était égal à leur taux de mutation. Néanmoins, au-delà de la possibilité de dater la divergence entre des séquences, l'hypothèse de l'horloge moléculaire décrivait également le processus évolutif des organismes vivants, c'est-à-dire que l'évolution moléculaire était, pour la plupart, due au remplacement stochastique et progressif d'allèles fonctionnellement équivalents, les modifications favorables étant donc rares et sans impact sur le taux global d'évolution. Or, cette conceptualisation du processus évolutif va, de fait, à l'encontre du concept de la sélection positive. En effet, elle implique que toute modification de l'horloge évolutive est due soit à une évolution adaptative à des contraintes (par exemple environnementales, plus ou moins stringentes), ou un changement de la taille de la population (dont nous avons déjà expliqué les implications en termes d'évolution). Cependant, l'hypothèse de l'horloge moléculaire stricte étant la plus simple à aborder, tant d'un point de vue conceptuel que mathématique, elle sera le point de départ de toute analyse et de tout développement en datation moléculaire, de la même façon qu'à partir du modèle de Jukes et Cantor ont été développés des modèles plus complexes de phylogénie. Ces modèles sont d'ailleurs utilisés pour la construction des arbres qui vont guider l'analyse en datation moléculaire.

Selon l'hypothèse de l'horloge moléculaire stricte, des séquences ayant divergé à partir d'un ancêtre commun doivent avoir accumulé le même nombre de substitutions au cours du temps, puisque leur taux d'évolution est identique. Cependant, lorsque comme dans le cas des virus, la séquence ancestrale demeure inconnue, il n'y a aucune possibilité de vérifier cette hypothèse. Tout comme dans le cas de l'enracinement des arbres phylogénétiques, l'incorporation dans l'analyse d'un ou de plusieurs groupes externes peut pallier cette absence puisque, dans le cas d'une horloge moléculaire stricte, le nombre de substitutions entre la séquence du (ou des) groupe(s) externe(s) et chaque taxon doit être identique. Néanmoins, cette méthode dite de « régression linéaire de la racine aux feuilles » (*root-to-tip linear regression method*) (Drummond *et al.* 2003a), même si elle a été développée pour permettre la comparaison d'un nombre assez important de taxa, n'atteste pas la possibilité de détecter un taux d'évolution variable (Bromham *et al.* 2000). En effet, l'écueil majeur auquel cette méthode se heurte est d'avoir à effectuer de nombreux tests sur des évènements non-indépendants, ce qui conduit souvent à surestimer le taux d'évolution. En effet, cette méthode repose sur l'hypothèse de l'indépendance des distances génétiques liant des paires de séquences, alors que dans les faits, les séquences sont liées par un ancêtre commun et une histoire évolutive partagée. La non indépendance des distances génétiques est un problème classique en analyse évolutive (Harvey & Pagel 1991), mais il peut être résolu par l'utilisation de méthodes prenant en compte de manière explicite

la structure implicite des données de séquences étudiées. Ainsi sera fait en appliquant aux données les modèles évolutifs que nous avons décrits précédemment.

Le maximum de vraisemblance permet la prise en compte d'un ensemble d'évènements non indépendants pour comparer des hypothèses, via le test LRT. Le test LRT utilise les valeurs de vraisemblance attribuées à chaque hypothèse, dont le double de la différence doit normalement suivre une loi du χ^2. Ce test a été appliqué à la détermination de l'horloge moléculaire, afin d'évaluer si le taux de substitution était homogène ou non, tout au long des branches d'un arbre phylogénétique. Pour y parvenir, le test LRT observe si la longueur des branches respecte l'horloge moléculaire stricte, c'est-à-dire si le taux de substitution instantanée moyen est identique entre chaque branche. Le test compare donc la longueur totale d'un arbre construit sans hypothèse d'horloge moléculaire (i.e. un arbre non raciné), ayant $2n - 3$ branches, à celle d'un arbre construit sous hypothèse d'horloge moléculaire (i.e. un arbre raciné) ayant $n - 1$ branches. Il est à noter que la différence dans le nombre de branches entre les deux arbres, provient du fait qu'un arbre construit avec l'horloge moléculaire est ultramétrique, c'est-à-dire que toutes ses feuilles sont équidistantes à la racine, puisque la quantité d'évolution dans un intervalle de temps donné est identique tout le long des branches. Ainsi, il n'est pas utile de calculer la longueur de chacune des branches, puisque, pour n'importe quel couple de taxa, la longueur des branches menant à leur ancêtre commun est identique. L'une peut donc être inférée de l'autre. Il est cependant important de noter que le test LRT peut conduire à rejeter l'hypothèse d'horloge moléculaire stricte en cas de présence de recombinaisons dans les séquences étudiées (Schierup & Hein 2000a, b). Il conviendra donc de vérifier l'absence de recombinaison dans le jeu de données avant de construire les arbres. Pour davantage de précisions sur le test LRT et l'horloge moléculaire, se rapporter à la partie résultats de ce manuscrit, dans laquelle il est détaillé.

Dans la plupart des cas, les analyses en phylogénie moléculaire faites sur des isolats viraux sont effectuées à partir de jeux de données de séquences étant issus d'un échantillonnage sériel, c'est-à-dire provenant de virus isolés à différents temps et en différents lieux. De ce fait, la date d'isolement est en général connue. Si le taux d'évolution est continu dans le temps, ces virus forment alors une population dont l'évolution est mesurable (Drummond *et al.* 2003b) (MEPs, pour *measurably evolving populations*), car les séquences provenant des isolats disponibles ont évolué de la même manière depuis leur ancêtre commun. On observe au contraire et principalement pour les virus à ARN, un taux d'évolution ayant tendance à s'accélérer au cours du temps. Quoiqu'il en soit, les analyses en datation moléculaire tiendront compte de la date d'échantillonnage. Elles ne pourront donc plus générer un arbre ultramétrique, puisque le calcul de la longueur des branches des feuilles à la racine, sera contraint par ces dates. De fait, la connaissance de l'âge des feuilles permettra d'inférer l'âge des nœuds internes de l'arbre, en le traduisant en un taux d'évolution par site et par unité de temps (Rambaut 2000) et ainsi, de calibrer l'horloge moléculaire réelle s'appliquant aux séquences considérées.

Cependant, la régularité de l'horloge moléculaire n'est pas aussi précise que le suppose la théorie (Bromham & Penny 2003). Ainsi, de nombreuses études ont montré des taux d'évolution variables (Jenkins et al. 2002) dont la non prise en compte conduirait le plus souvent à une datation erronée et au masquage d'un taux de substitution possiblement « lignée-dépendant ». En effet, si selon la théorie la fréquence des substitutions survenant dans une séquence suit une distribution de Poisson (Zheng 2001), l'horloge moléculaire est en fait souvent moins rigoureuse, conduisant la fréquence des substitutions à suivre une distribution de Poisson « élargie » (Cutler 2000). Bien des facteurs peuvent être à l'origine d'une variation dans le taux de substitution, ne serait-ce que la modification du rapport entre la dérive génétique et la force évolutive sélectionniste. Cette dernière, peut entrainer une forte poussée évolutive en raison des contraintes imposées à la protéine en raison d'un changement de l'environnement ou de la pression du système immunitaire obligeant par exemple un virus à muter pour survivre. De même, le temps de génération ou encore les mécanismes de réparation de l'ADN à l'œuvre au cours de sa réplication, peuvent aussi avoir une influence sur le taux d'évolution d'une séquence (Bromham & Penny 2003). Ces contraintes évolutives pouvant varier selon les époques, les environnements ou les hôtes, l'effet « lignée-dépendant » peut s'avérer être un facteur déterminant pour le taux d'évolution. D'ailleurs, Drake a clairement mis en lumière cette variation du taux de substitution entre les taxa (Drake *et al.* 1998). Ce relâchement dans la régularité de l'horloge aura donc pour conséquence une datation à tout le moins approximative, voire erronée. Pour circonvenir ce problème, des méthodes de datation moléculaire ont été développées, qui permettent de modéliser un taux de substitution variable en un processus suivant tout de même une distribution de Poisson (Felsenstein 1981 ; Rambaut & Bromham 1998). Néanmoins dans ces modèles d'horloge relâchée, il n'a pas été possible de caractériser les paramètres comme cela a été le cas pour les modèles évolutifs que nous avons détaillés dans le chapitre précédent.

L'un des premiers modèles à avoir été développé permettait à un ou plusieurs clades d'avoir un taux d'évolution constant différant des autres (Yoder & Yang 2000). Cette méthode a pris le nom d'horloge moléculaire « locale ». Le fait d'assigner à un clade particulier un taux propre d'évolution présuppose que l'arbre phylogénétique qui détermine les différents phyla ait été préalablement construit de façon optimale, afin de représenter le plus fidèlement la réalité des liens unissant les séquences étudiées. En effet, si les relations entre taxa sont mal résolues, la délimitation des clades sera approximative et pourra conduire à regrouper des taxa pourtant divergeant, assignant un taux de substitution identique à des séquences qui n'évoluent pas réellement au même rythme. Cette méthode permet néanmoins de comparer différents arbres entre eux, via le test LRT (Kumar & Hedges 1998 ; Takezaki et al. 1995). Cependant, le test LRT fixe l'hypothèse nulle (ici, l'arbre construit selon un modèle d'horloge moléculaire locale) pour la comparer avec l'arbre non raciné, c'est-à-dire construit sans horloge moléculaire. Or, il s'avère que, dans ce cas, la valeur du χ^2 résultant du test LRT peut ne pas être suffisamment significative, et conduire à ne pas apprécier correctement la différence entre les arbres, tout particulièrement lorsque

les séquences utilisées sont courtes. Ainsi, les taux de substitution variables seraient difficilement détectables avec ces méthodes, ce qui conduirait à dater l'ancêtre commun des séquences beaucoup trop loin dans le temps (Bromham *et al.* 2000).

Comme dans le cas des reconstructions phylogénétiques, une approche bayésienne de la datation moléculaire a été développée (Kishino *et al.* 2001 ; Sanderson 1997 ; Thorne *et al.* 1998). Les modèles bayésiens ont prouvé leur justesse en réconciliant, par exemple, le taux d'évolution et les données paléontologiques des lignées animales ancestrales (Aris-Brosou & Yang 2002). Les analyses bayésiennes appliquées à la datation moléculaire utilisant une horloge relâchée, permettent de spécifier *a priori* que le taux de substitution le long d'une branche de l'arbre suit une distribution lognormale, centrée sur le taux de substitution des branches ancestrales. De cette façon, les branches ascendantes et descendantes sont corrélées. Le taux de chaque branche est alors déterminé à partir d'une distribution paramétrique, dont la moyenne est une fonction du taux d'évolution de la branche parentale. Il est aussi possible d'envisager une distribution exponentielle, ce qui impliquerait que les variations de taux d'évolution n'apparaissent plus le long des branches, mais au niveau des nœuds, sans lien avec la longueur des branches. Cependant, ces modèles auto-corrélés présupposent que des lignées proches auront un comportement évolutif également proche, ce qui n'est pas absolument avéré. En effet, l'auto-corrélation des séquences signifie que la majorité des différences observées est due à la filiation. Or, lorsque l'on réduit l'échelle de temps séparant l'apparition de deux séquences, les variations purement stochastiques et les contraintes environnementales semblent plus importantes que la filiation pour expliquer les différences observées. D'un autre côté, la seule filiation, pour expliquer les différences de taux d'évolution, semble très improbable lorsque l'on augmente l'échelle de temps ou de distance entre deux échantillonnages. La difficulté réside donc dans le positionnement d'une frontière entre la filiation et les autres facteurs engendrant la variation du taux d'évolution entre des séquences, c'est-à-dire dans l'expression des limites de leur auto-corrélation.

Des alternatives à ces modèles ont donc alors été développées, dans lesquelles les branches adjacentes d'un arbre ne sont pas corrélées (Drummond *et al.* 2006). Ces modèles d'horloges moléculaires non corrélées permettent de déterminer le taux d'évolution de chaque branche, ou de chaque nœud (selon que la distribution *a priori* des taux soit envisagée comme lognormale ou exponentielle). L'avantage majeur de ces méthodes est qu'elles ne requièrent pas l'implémentation préalable de la topologie de l'arbre pour calculer les taux d'évolution, mais déterminent elles-mêmes le meilleur arbre en comparant puis en approximant les arbres générés au fil des MCMC (Drummond *et al.* 2006). Toutes les méthodes déterminant l'âge de l'ancêtre commun d'un jeu de données de séquences peuvent être soumises à discussion. La meilleure façon de confirmer ces hypothèses reste donc de pouvoir les étayer par des données historiques.

4- Les réseaux

Les biologistes, pour modéliser la biodiversité, ont traditionnellement utilisés des arbres, censés représenter les liens qui unissent les organismes dont ils étudient l'évolution. La démarche suivie pour démêler la complexité de ces relations a toujours été de déterminer d'abord le processus le plus simple pouvant l'expliquer, puis de le complexifier progressivement afin que sa valeur descriptive et/ou prédictive devienne de plus en plus fidèle à la réalité. Dans les arbres qui décrivent les relations entre les organismes, les nœuds internes représentent les ascendants inférés et les feuilles, les taxa réellement observés. La longueur des branches reliant les nœuds entre eux est relative à la quantité de variation entre les taxa (inférés ou observés). Ainsi, un arbre peut-il être interprété biologiquement pour démontrer une histoire évolutive. L'utilisation de tels arbres présuppose néanmoins que le mécanisme sous-jacent à l'évolution des taxa est un mécanisme strictement bifurcatif, c'est-à-dire qu'un descendant a toujours un seul ascendant. Or, cette ascendance unique ne modélise pas explicitement la réalité évolutive de tous les jeux de données de séquences. Dans le cas des phénomènes d'explosion évolutive, comme par exemple la primo-infection au virus HIV (Buendia & Narasimhan 2009), l'arbre inféré des données devrait présenter des polytomies pour rendre compte des véritables liens phylogénétiques entre les taxa. Ces polytomies figurent des incertitudes quant au patron réel des branchements de l'arbre, ce qui revient à dire que plusieurs topologies sont possibles pour résoudre le jeu de données. Dans le cas d'histoires évolutives complexes, les nœuds internes peuvent non seulement être inférés, mais être également des taxa réellement observés. Ainsi, la représentation des interactions entre les taxa, pour être fidèle à la réalité évolutive, devrait être modélisée en un réseau, c'est-à-dire un arbre réticulé.

Les réseaux ne présupposent pas un modèle évolutif en forme d'arbre et ne vont pas contraindre la résolution d'un jeu de données à devenir un arbre. Leur utilisation peut donc *a minima* indiquer si le meilleur modèle pour résoudre un jeu de données est ou non un arbre. Les réseaux peuvent montrer les histoires évolutives parallèles, mais non identiques, de lignées dérivant pourtant d'un ancêtre commun. Un réseau peut donc avoir deux interprétations biologiques possibles : (i) la représentation de liens non dichotomiques entre ascendants et descendants et (ii) la représentation de patrons évolutifs incompatibles résultant de conflits ou d'incertitudes dans la résolution des données (Morrison 2005). Dans le cas des populations dont la reproduction est sexuée par exemple, les liens entre organismes ne sont explicables que par un réseau, pour la simple raison que la reproduction sexuée implique l'inter-fécondation, ce qui génère un réseau de connections entre parents et descendants, appelé généalogie réticulée. La reproduction sexuée n'est cependant pas le seul mécanisme introduisant ces phénomènes de réticulation. En effet, dans le cas de mécanismes tels que les recombinaisons, les hybridations, les transferts latéraux de gènes ou les réassortiments géniques, les descendants ont manifestement également plusieurs ascendants (Arenas *et al.* 2008 ; Posada & Crandall 2001).

Les conflits entre des patrons de caractères différents peuvent rendre un arbre phylogénétique instable, ces conflits n'étant pas compatibles avec une représentation en un arbre unique. Les conflits peuvent avoir plusieurs origines : (i) des incertitudes ou des ambigüités dues à des données non suffisamment précises (défaut d'alignement des séquences, utilisation d'un modèle évolutif non adapté…), (ii) l'existence d'homoplasies dans les données, c'est-à-dire que les caractères, au lieu d'être hérités, proviennent de convergence (similarité des caractères, mais non réelle homologie) ou de reversions et (iii) d'évènements évolutifs impliquant des échanges de gènes ou de fragment de gènes entre des organismes différents.

La plupart du temps, ces conflits sont résolus en affectant un poids aux nœuds de l'arbre, comme c'est par exemple le cas avec le ré-échantillonnage des données (*bootstraps*). Une valeur faible de bootstrap signifie qu'il existe de nombreux conflits entre patrons de caractères, sans pour autant en expliciter la nature et la localisation. En effet, l'arbre phylogénétique produit est l'arbre majoritaire, les arbres conflictuels représentant les patrons mineurs restant discrets. Les interconnections des branches, dans un réseau vont permettre de spécifier la nature et la localisation des conflits. Ainsi, les parties du réseau ressemblant à un arbre dichotomique vont représenter une absence de conflit entre les patrons de caractères, tandis que les multifurcations attesteront un manque de données pour résoudre leur phylogénie. Enfin, les réticulations (également appelées anastomoses) montrent l'existence d'un conflit entre au moins deux patrons de caractères. Ainsi, le réseau montre visuellement les arbres alternatifs compatibles avec le jeu de données. Tout comme dans le cas d'un arbre bifurcatif, la distance phénétique entre deux taxa est proportionnelle au nombre de changements d'état des caractères ou à la distance génétique à la différence près qu'il peut exister plusieurs chemins pour la représenter.

La façon de construire les réseaux est très semblable à celle utilisée pour construire les arbres. Concernant les arbres, il s'agit de regarder les états des caractères dans le jeu de données, puis de les résumer (les « afficher ») en un diagramme sur lequel on place une racine afin de lui donner une direction évolutive. Il n'existe ainsi qu'un chemin unique et sans ambigüité entre la racine et chaque nœud ou feuille de l'arbre, chaque nœud interne inférant un ancêtre. Ainsi, les caractères et leur interprétation phylogénétique sont-ils réunis sur un seul et même diagramme. Cette possibilité provient du fait que les méthodes de construction d'arbres délaissent les ambigüités (rares) qui existent dans les données pour ne s'intéresser qu'aux évènements évolutifs « vrais » (plus nombreux). La construction d'un réseau selon cette approche « d'affichage des caractères » reviendrait à superposer les arbres correspondant à tous les patrons possibles, et compatibles avec les données, les plexus anastomotiques représentant les zones conflictuelles. L'enracinement du réseau permettra de lui donner une direction évolutive. Cependant un tel réseau n'est pas interprétable sur le plan de l'évolution de l'état des caractères puisque l'ensemble des nœuds, tout comme l'ensemble des réticulations dans le réseau ne correspondent pas obligatoirement à un ancêtre (Nakhleh *et al.* 2003), ni à un évènement évolutif. En tant que

résumé mathématiques d'arbres multiples (Bryant 2003), les réseaux permettent donc d'explorer les données, mais ne résolvent pas leur véritable phylogénie.

Pour construire des réseaux phylogénétiques explicites, des méthodes ont été développées visant à modéliser les processus sous-jacents aux réticulations des arbres phylogénétiques (recombinaisons, hybridation, transferts latéraux de gènes…) en détectant les patrons de caractères que ces processus induisent dans les données. Le principe de construction assume un arbre raciné auquel on ajoute des réticulations selon un modèle mathématique déterminé en fonction du processus biologique choisi. Les séquences ancestrales peuvent alors évoluer en séquences descendantes, le réseau étant contraint par une direction temporelle due à la racine. Plusieurs types de réseaux phylogénétiques explicites peuvent être construits selon cinq différentes méthodes.

Les réticulogrammes sont des réseaux construits selon une méthode basée sur les distances (Legendre & Makarenkov 2002 ; Makarenkov & Legendre 2004). Un premier arbre est d'abord construit selon une méthode de distance, puis les réticulations sont ajoutées progressivement de façon à optimiser des critères d'adéquation du réseau aux données. Le principe sous-jacent à cette méthode est que les phénomènes de réticulation sont plus rares que les dichotomies dans un jeu de données et donc le réseau ne devrait pas trop dévier de l'arbre phylogénétique qui reste le modèle le plus simple pour représenter les données. Les réticulations, qui représentent donc les différents patrons de caractères non explicites dans l'arbre original, sont parfois difficilement interprétables. Comme toutes les branches sont représentées dans ces réseaux, y compris celles ayant une longueur nulle, le diagramme est d'autant plus difficile à lire lorsque le nombre de caractères non résolus croît. De plus, malgré la racine, la direction des branches peut rester ambigüe.

La méthode statistique en parcimonie permet également de construire des réseaux (Templeton *et al.* 1992). Le principe est alors de connecter les taxa en fonction de l'augmentation des différences de caractères observés entre eux, dans la limite de la fiabilité phylogénétique parcimonieuse car la parcimonie est sujette au phénomène d'attraction des longues branches. Comme en parcimonie, chaque branche représente un changement particulier de caractère. L'interprétation biologique est la même que pour un arbre parcimonieux, puisque les nœuds ancestraux sont explicitement inférés. Cette méthode peut néanmoins aboutir à la construction de réseaux multiples et disjoints lorsque la diversité des caractères est trop élevée, et la direction des branches même en présence d'une racine, reste parfois ambigüe.

Des méthodes basées sur les caractères ont aussi été développées, permettant de construire des réseaux dits médians, principalement appliquées aux caractères binaires (Bandelt *et al.* 2000). Ces méthodes affichent visuellement tous les états des caractères afin de voir les incompatibilités entre les différents patrons, mais demande une nouvelle dimension pour chaque nouveau patron implémenté. Le diagramme devient donc vite très complexe avec l'augmentation du nombre de différences entre les caractères.

L'interprétation biologique est possible, les branches parallèles étant contraintes par le modèle à être unidirectionnelles et sans ambigüité en présence d'une racine. Toutefois, les nœuds internes du réseau ne représentant pas tous un ancêtre inféré, interpréter toutes les réticulations en tant qu'évènement évolutif augmente le risque de génération de faux positifs.

Une quatrième famille de méthode a été développée, utilisant le principe de la « décomposition par partitionnement » (*split decomposition*). L'objectif de cette méthode est d'afficher le maximum d'états des caractères dans un diagramme à deux dimensions. Le diagramme peut être construit en utilisant la parcimonie, mais le plus souvent la méthode utilise une mesure des distances entre les caractères (Bandelt & Dress 1992b ; Bandelt 1992a). Le réseau inféré représente une collection de différentes bipartitions A et B non vides d'un jeu de données D, de telle manière que $A \cup B = D$ et que $A \cap B = \emptyset$. Pour chaque partition du jeu de données, une distance est définie et un index d'isolement est calculé en utilisant des quartets de partitions. Cet index établit si la distance calculée peut soutenir la partition réalisée. La décomposition par partitionnement utilisera alors les distances affectées à chaque partition ainsi que l'index d'isolement afin de trouver la réalisation minimale permettant de résoudre le jeu de données. En effet, les relations unissant quatre taxa peuvent être décrites en construisant trois arbres non enracinés. Si un seul taxon diffère par ses caractères, un seul arbre sera construit, tandis que si au moins deux taxa diffèrent, ce sera un réseau. L'affichage des conflits dans ces réseaux génère des multifurcations non informatives (représentant des faux négatifs) lorsque le conflit est trop complexe. De plus, les nœuds internes ne représentant pas tous des ancêtres, l'interprétation biologique peut être rendue difficile, bien que les branches parallèles soient unidirectionnelles.

Nous venons de le voir, les réseaux médians génèrent des faux positifs, tandis que la décomposition par partitionnement génère des faux négatifs. Pour pallier ces deux problèmes, la méthode dite des « réseaux voisins » (*neighbor-net*) a été développée (Bryant & Moulton 2004). Basée sur l'utilisation des distances entre les taxa, cette méthode tend à généraliser les méthodes du plus proche voisin (*neighbor-joining*) en affichant les données en deux dimensions. C'est la meilleure méthode pour résoudre les phylogénies complexes, même si son interprétation biologique est parfois difficile. En effet, bien que la racine et le modèle infèrent des branches parallèles unidirectionnelles, certains des nœuds internes ne correspondront pas à des ancêtres.

5- Les virus dans l'histoire évolutive du vivant

5-1- L'origine des virus

La détermination par Watson et Crick, de la structure de la molécule d'ADN, support de l'information génétique à l'origine du phénotype de tout organisme a mis en évidence l'universalité du code génétique (à quelques très rares exceptions près). Ces découvertes fondamentales ont eu pour conséquence d'unifier le vivant : le vivant n'était dès lors plus pluriel, car il utilisait la même machinerie de synthèse des macromolécules qui le compose, et à partir du même code génétique. Il était donc « seulement » divergeant. Or, qui dit divergence dit origine commune à partir de laquelle les divergences s'accumulent au cours du temps, jusqu'à la séparation des lignées ayant trop divergées par rapport à leur ancêtre commun. Ainsi, de spéciation en spéciation, la diversité observée peut être expliquée. La question qui se pose est alors de remonter à l'ancêtre commun de tous les êtres vivants, sans présomption de sa nature intrinsèque. C'est ainsi que le concept du plus proche ancêtre commun unique (LUCA) a été introduit (cf. supra). LUCA est donc notre dernier ancêtre commun, c'est-à-dire, non pas la première hypothétique cellule ayant vécu sur Terre, mais le plus proche de nos ancêtres dans le temps à partir duquel auront émergé les trois grands règnes actuellement déterminés : eucaryotes, eubactéries et archées. S'il reste cependant difficile de déterminer avec précision la nature même de LUCA, les analyses en évolution moléculaire ne peuvent le supposer qu'unique. L'unicité de cet ancêtre commun n'est pas un simple mythe. La génomique comparative, grâce à l'étude des homologies a en effet mis en évidence l'existence de protéines « universelles », présentes chez tous les êtres vivants, donc également constitutives de LUCA. Leur nombre est évalué à environ un peu moins d'une centaine (Koonin 2003). Les êtres vivants partageant tous la même machinerie de synthèse de leurs protéines, il n'est pas étonnant que la plupart des protéines universelles soient des enzymes impliquées dans la traduction des protéines, preuve que le vivant tel que nous le connaissons aujourd'hui, procède et a toujours procédé du même processus pour se maintenir et se reproduire. Le fait que, parmi ces protéines universelles, se trouvent également des protéines de translocation membranaire des protéines néo-synthétisées, laisse à penser que LUCA était lui-même un organisme cellulaire, séparé de son environnement par une membrane (Pereto 2004). Comment concevoir de toute façon le développement de métabolismes élaborés tels que ceux qui se sont établis dans les cellules, en l'absence du confinement cellulaire ? Il est même quasiment établi que la cellularisation du vivant a eu lieu bien avant l'émergence de LUCA (Jeffares *et al.* 1998). La question de savoir si LUCA possédait un génome composé d'ADN ou d'ARN reste encore non définitivement tranchée. Cependant, des études ont montré qu'il existait des mécanismes de réplication de l'ARN permettant de détecter et de réparer les erreurs de duplication : cette absence de démonstration était jusqu'alors l'argument principal des tenants d'un LUCA à ADN génomique (Poole 2005).

Qu'en est-il des virus et quels pourraient être leurs liens avec LUCA ?

Au cours de l'évolution, des bactéries ancestrales (également appelées α-protéobactéries), sont devenues symbiotiques dans les cellules eucaryotes et ont donné les mitochondries. Les polymérases ADN et ARN ainsi que la primase (une enzyme très impliquées au niveau de la fourche de réplication de l'ADN) de la bactérie ancestrale ont été remplacées par une ADN polymérase, une hélicase et une primase tirant leur origine d'un virus de parenté commune avec les bactériophages T3 et T7 (Filee 2003). Cette observation présuppose l'antériorité des virus sur celle de LUCA. Certains estiment que LUCA pourrait avoir été une cellule à ARN et que l'ADN aurait été inventé par des virus pour se prémunir des mécanismes de défense cellulaire (Forterre 2002). Cette invention aurait ensuite été transmise à leurs hôtes. Ainsi, les virus auraient apporté aux organismes cellulaires qu'ils parasitent des caractéristiques nouvelles telle que la capacité à synthétiser de l'ADN à la place de l'ARN au moyen d'une retro-transcriptase (Forterre & Gadelle 2009). Cette théorie possède ses détracteurs, estimant que les virus ont évolué en utilisant et en incorporant dans leur propre génome certains gènes de la machinerie cellulaire, gènes utiles à leur descendance, comme par exemple la retro-transcriptase du retron archéobactérien (Becker 1998).

Quoiqu'il en soit, il semble désormais communément accepté que la molécule d'ARN a préexisté à celle de l'ADN (Ribas de Pouplana 2004), ne serait-ce que parce que la molécule d'ADN peut être considérée comme une molécule d'ARN modifiée, le ribose étant le sucre « naturel » comparé au désoxyribose et la thymine simplement une uracile méthylée (Forterre 2005). Pour autant, le vivant peut-il être réduit à une simple molécule d'ARN ? Certains ont considéré que dans le monde pré-biotique, les molécules d'ARN étaient libres (Gilbert 1986). Cependant, il semble improbable que des molécules d'ARN libres dans leur environnement soient parvenues à développer des mécanismes tels que les complexes ribosomaux, capables de traduire de l'information génétique en protéines. De plus, la nature non cellulaire de LUCA serait en contradiction avec l'observation des protéines membranaires universelles, telles que les enzymes impliquées dans la biosynthèse lipidique (Pereto 2004). LUCA n'était vraisemblablement pas une cellule à génome d'ARN, ces cellules devant avoir été éliminées au cours de l'évolution par le processus darwinien puisque l'ADN est nettement plus stable que l'ARN (notamment à cause de la grande réactivité de l'atome d'oxygène en 2' du ribose de l'ARN). Ces cellules à ARN ne pouvant représenter l'ancêtre du monde vivant moderne, ont cependant donné naissance au cours de l'évolution, aux cellules à ADN, et c'est celui-ci qui est parvenu jusqu'à nous. Il est d'ailleurs tout aussi probable que les virus à ARN aient été préexistants aux virus à ADN.

Reste la question de la transition des génomes à ARN en génomes à ADN. L'ADN est une molécule beaucoup plus stable que celle d'ARN ce qui lui confère un avantage sélectif au sens darwinien du terme pour la conservation et la transmission de l'information génétique nécessaire à la fabrication des protéines. Il est aujourd'hui avéré que certains virus à ARN se

sont transformés en virus à ADN, précisément dans le but d'échapper aux mécanismes de destruction de l'ARN (Warren 1980). Cette transformation leur aurait ainsi permis d'échapper à tous les mécanismes de défense mis en place par les cellules à génome ribonucléique. Or, les mécanismes d'échappement aux défenses de l'hôte par les virus sont tellement variés, comme nous le verrons concernant le virus de la PPA, que la simple modification de l'ARN en U-ADN puis en T-ADN doit avoir été un mécanisme relativement simple à mettre en place au cours du temps long de l'évolution. En fin de compte, la transition entre virus à ARN et virus à ADN n'aura nécessité que la présence d'une enzyme de retro-transcription, puis l'apparition d'une enzyme ADN-polymérase ADN-dépendante. Cette propriété de transmutation de l'ARN en ADN aura ensuite été transmise aux cellules parasitées, de la même manière qu'un plasmide peut conférer de nouvelles caractéristiques à la bactérie qui les intègre. On trouve d'ailleurs de nombreuses homologies entre les protéines codées par les plasmides et celles codées par les virus (Forterre 2004), suggérant là encore, une origine commune suivie d'une coévolution.

De notre point de vue, les virus seraient donc les formes résiduelles (ou restreintes) ultimes de la coévolution de toutes les formes cellulaires étant apparues au cours de l'évolution. Ils sont d'ailleurs les seuls organismes exploitant encore tous les mécanismes réplicatifs imaginés par le vivant pour se transmettre à sa descendance, même si pour ce faire ils doivent « emprunter » aux cellules qu'ils parasitent la plus grande part de la machinerie nécessaire à ces mécanismes. Il est à cet égard remarquable de constater que seuls les virus disposent désormais d'enzymes telles que les retro-transcriptases (à l'exception des rétrotransposons qui en disposent eux aussi) et les polymérases à ARN dépendantes de l'ARN. Les virus ancestraux auraient donc de tout temps parasité tous les types cellulaires et leur taux élevé d'évolution aurait permis la génération d'une multitude de gènes que se seraient appropriés les organismes cellulaires ancestraux lorsqu'ils leur conféraient un avantage sélectif. Chaque organisme des trois règnes du vivant actuel héberge des virus en son sein, suggérant là encore une coévolution (car la plupart de ces virus ne sont ni lytiques, ni lysogéniques et correspondent donc plus à des symbiontes qu'à des parasites pathogènes). Ils possèdent encore tous les types de supports connus de l'information génétique : ARN et ADN, simple et double brins, segmentés ou non, circulaires ou linéaires. Or, de tous ces supports, seul l'ADN a été conservé par le vivant au cours de l'évolution. LUCA serait donc, en fin de compte, une cellule à ARN qui aurait acquis des virus la capacité de transformer son ARN en ADN. Pour expliquer les trois règnes du vivant, LUCA aurait pu recevoir cette capacité de la part de trois virus différents, dont les transformations engendrées chez LUCA auraient *in fine* conduit à la différenciation en eucaryotes, eubactéries et archées (Forterre 2006). Les virus sont ainsi souvent considérés comme des « accélérateurs évolutionnaires » (Bubanovic *et al.* 2005).

Le passage spéculatif des organismes à ARN aux organismes à ADN représente une véritable orientation prise par le vivant, les virus pouvant donc être considérés comme des fossiles vivants (?), une sorte de témoignage de l'ensemble des balbutiements directionnels

des origines. Certains évoquent même la possibilité que les rétrovirus et les hepadnavirus, de par leur capacité à manipuler les deux types de molécules, pour passer de l'une à l'autre, représentent la forme intermédiaire symbolique de ce passage (Miller & Robinson 1986). Cependant, même s'ils sont soumis à des forces évolutives identiques, les natures biochimiques et structurelles différentes des génomes des virus vont induire inévitablement un équilibre différent des mécanismes évolutifs, et donc un rythme évolutif dissemblable.

5-2- L'évolution des virus

Les substitutions nucléotidiques représentent la base de toute évolution des organismes. Les forces évolutives telles que la dérive génétique, la dérive antigénique, les recombinaisons, etc… vont jouer un grand rôle dans le façonnage de la structure génétique des individus. Les virus, en tant que parasites intracellulaires obligatoires sont particulièrement soumis à ces forces évolutives, ne serait-ce que pour échapper à la réponse des hôtes qu'ils infectent.

Malgré une littérature de plus en plus abondante concernant les processus mutationnels à l'œuvre chez les organismes vivants, peu d'espèces ont vu leur taux de mutation, ni les déterminants évolutifs qui les induisent, étudiés. Drake a été l'un des premiers à s'intéresser au taux de mutations spontanées apparaissant à chaque réplication des génomes d'un organisme (Drake *et al.* 1998). Le taux de mutation correspond au taux de variations nucléotidiques apparaissant dans un génome lors de sa réplication. Il est à distinguer du taux de substitution qui définit le nombre de mutations fixées au sein d'une population, c'est-à-dire transmises à la descendance. Ces deux notions sont intimement liées et si les processus évolutifs à l'œuvre sur une séquence d'ADN sont neutres (c'est-à-dire n'induisant pas de transformation dans les protéines traduites), la relation entre ces deux phénomènes est simple (Li 1997). En revanche, si les mutations sont non synonymes, l'étude du taux et de la dynamique de l'apparition des mutations permettra de définir si les mécanismes sous-jacents à ces mutations relève d'un processus sélectif, c'est-à-dire non purement stochastique.

Drake, s'est tout d'abord intéressé au taux de mutation dans les génomes de virus lytiques à ARN (Drake 1993). Il a observé le taux de mutation par base (μ_b) dont il a ensuite déduit le taux de mutation par génome et par réplication (μ_g). Par exemple, il a constaté que le taux μ_b du bactériophage Qβ était de 1,5 x 10^{-3} (soit un taux $\mu_g = 6{,}5$ pour un génome de 4215 nucléotides), tandis que chez le virus *influenza A*, μ_b était > 7,3 X 10^{-5} pour un $\mu_g \geq 1$ (génome d'environ 15000 nucléotides). La réplication des génomes de virus lytiques à ARN se faisant en continu durant l'infection, le nombre de particules virales contenues dans une cellule infectée et dont le génome a muté peut devenir très élevé (dans le cas du virus influenza, un mutant pour 10 virus produits). De fait, une population virale

issue d'un virus unique est hétérogène. On parle de quasi-espèces. Il est à noter qu'un taux de mutagénèse élevé chez les virus à ARN aboutit dans la plupart des cas à des conséquences létales pour les sous-populations (ou quasi-espèces) (Eigen 2002, 1979). Ce taux d'erreurs a d'ailleurs été qualifié par Eigen « *d'erreur catastrophique* ». Ce taux de mutation élevé permet une dynamique réplicative à l'intérieur du spectre des populations de mutants. Si ces populations coopèrent, il y aura une augmentation de la capacité réplicative des virus, tandis que si les mutants interfèrent les uns avec les autres les fonctions virales se verront détériorées, jusqu'à être létales pour la population entière (Perales *et al.* 2012). Ce concept de taux d'erreurs létal est d'ailleurs à l'origine de nombreuses théories thérapeutiques anti-virales (Ochoa 2006).

Le taux de mutation des virus à ADN est très inférieur à celui des virus à ARN. Ainsi, chez le bactériophage M13, μ_b = 7,2 x 10^{-7} (pour un μ_g = 0,0046), soit un taux mille fois inférieur à ceux des virus à ARN. Les rétrovirus ont quant à eux une position intermédiaire. A titre de comparaison, les taux μ_b des eucaryotes, tels que la drosophile ou la souris sont de l'ordre de 10^{-10}, et descend jusqu'à 10^{-11} chez l'homme. Cependant, la taille importante des génomes de ces organismes rendent des taux μ_g relativement élevés (0,16 pour l'homme) (Drake *et al.* 1998). Si ce taux de mutations par réplication dans le génome entier semble élevé, la présence de nombreux introns (parties du génome non codantes) chez les eucaryotes et qui sont souvent des loci à haute fréquence mutationnelle, explique aisément qu'elles ne soient pas létales pour ces organismes. Chez les procaryotes, en revanche, l'ensemble du génome (ou sa quasi-totalité) de taille beaucoup plus restreinte, est codant. L'impact des erreurs lors de la réplication est alors très augmenté. Une des explications dans les différences observées de taux de mutation entre les virus à ARN et les virus à ADN tient sans doute de la fidélité des polymérases qui sont utilisées au cours de leur cycle réplicatif respectif. Des études ont montré que le taux d'erreurs induites par les polymérases ARN dépendant de l'ARN et les polymérases ARN dépendant de l'ADN était nettement plus élevé que le taux d'erreurs induit par les polymérases ADN dépendant de l'ADN, probablement en raison de l'existence de mécanismes de détection et de réparation des erreurs de réplication par les complexes ADN-polymérases ADN-dépendants (Flint 2004 ; Garcia-Diaz 2007). Bien que des mécanismes de détection et de réparation de l'ARN existent durant sa réplication (Poole 2005), les mésappariements entre nucléotides ne peuvent être aussi fidèlement réparés dans les molécules doubles brins d'ARN, ou les hétéroduplexes ARN/ADN (Garcia-Diaz 2007).

Depuis les premières études menées par Drake, les capacités de séquençage ont été considérablement augmentées, permettant désormais des études à l'échelle de génomes entiers d'organismes, plutôt que de petits loci (Clark *et al.* 2007). Ceci permet d'appréhender la dynamique d'évolution des organismes dans leur entièreté, plutôt que par leurs gènes, dont l'histoire évolutive peut être différente de l'organisme entier, en raison par exemple de phénomènes de recombinaisons ou de transferts latéraux de gènes. Les organismes simples comme les virus, ont largement bénéficié de ces nouvelles capacités d'études et d'analyses,

et ce pour plusieurs raisons : (i) étudier le taux d'évolution chez des organismes qui évoluent beaucoup est plus démonstratif, et (ii) les temps de génération des virus sont extrêmement brefs, ce qui permet l'apparition de nombreuses mutations et donc d'une diversité génétique facilement observable. Ce sont donc principalement les virus à ARN qui ont été étudiés, puisque ce sont eux qui répondent le mieux à ces deux caractéristiques. Ils possèdent en outre l'avantage que leur génome est de petite taille, leur cycle réplicatif souvent bien connu, et leurs structures et fonctions largement documentées. Il était donc possible, au travers de l'étude de leur génome, de montrer les implications de leur évolution génomique sur le plan structurel et physiologique. Les virus à ADN à simple ou double brins n'ont pour autant pas été totalement oubliés (Drake *et al.* 2005).

La divergence entre les virus d'une même famille, ou d'un même genre, est en général basée sur l'étude de séquences provenant d'un échantillonnage dit sériel, c'est-à-dire provenant de virus qui ont été isolés au cours du temps, et dans des lieux généralement différents. Les divergences accumulées constatées sont donc la conséquence du phénomène de taux de substitution, c'est-à-dire, comme énoncé plus haut, de la fixation des mutations dans les descendances. Ce taux de substitution ne peut donc pas être linéaire dans le temps, et correspond donc à l'observation d'un état, et non d'un mécanisme évolutif observé à partir d'un clone. L'observation d'un tel mécanisme ne peut être réalisée que lorsque l'ensemble de la population étudiée a pour origine l'infection d'une cellule par un seul virus. Si l'ensemble de la descendance possède un parent unique, le génome parental (phénomène dit de « *stamping machine* »), les accumulations de mutations peuvent alors prendre une allure de progression à caractère linéaire. Ainsi, le taux de mutation peut être déterminé, en plus du taux de substitution, au sein de cette population. Cependant, même dans le cas d'une infection première par un unique virus, les premières molécules d'acides nucléiques filles produites peuvent à leur tour servir de matrice à la réplication et la production de nouveau virions. Dans ce cas, la progression de l'accumulation des mutations ne sera pas linéaire, mais géométrique (French 2003). Quel que soit le cas envisagé, la détermination des taux de mutation et de substitution dans ces populations permet non seulement d'étudier la diversité observable entre les organismes, accumulée au cours du temps, mais également de comparer entre elles des populations différentes d'individus (Drummond *et al.* 2003b).

Les divergences observées entre les séquences d'ARN ou d'ADN de virus proviennent des modifications par les enzymes réplicatives cellulaires, mais également au cours de la réplication, de l'apparition de recombinaisons entre les génomes dupliqués, pouvant conduire à des indels (Domingo 1997 ; Domingo & Holland 1997 ; Gibbs 1995). Le taux de mutations le plus élevé a été déterminé pour un virus à ARN simple brin, le bactériophage Qβ, comme nous l'avons vu plus haut ($1{,}5 \times 10^{-3}$ substitutions par site et par réplication) et la plus basse pour un virus à ADN double brin chez le virus HSV1 (*herpes simplex virus* type 1) et évaluée à $1{,}8 \times 10^{-8}$ (Drake & Hwang 2005). Ces taux correspondent peu ou prou aux taux d'erreur, ou de fiabilité, des polymérases virales mises en œuvre (ARN-polymérases ARN ou ADN-dépendantes, rétro-transcriptases, ADN-polymérases ADN-dépendantes). Au final, le

taux de mutation des virus peut être schématisé comme suis : $\mu_b$virus à ARN > μ_b rétrovirus > μ_b virus à ADN. Néanmoins, certains virus à ADN évoluent plus rapidement qu'attendu, tandis que certain virus à ARN, évoluent plus lentement. D'autres déterminants que les erreurs dues aux enzymes de polymérisation du matériel génomique doivent donc intervenir. Par exemple, des virus intégratifs comme le HIV ou le virus de l'hépatite B, lorsqu'ils sont libres dans le cytoplasme cellulaire ont un taux de mutation par site et par cycle réplicatif compris entre 0,1 et 0,2, ce qui reste faible comparé au virus utilisant des ARN-polymérases ARN-dépendantes. On observe cependant chez ces virus une dynamique évolutive très forte. Ces dynamiques reflètent donc majoritairement des taux de recombinaisons élevés, couplés à une sélection positive très forte (Bonhoeffer 1995 ; Drake 1993 ; Drake *et al.* 1998), ce qui affecte grandement et positivement leur capacité à évoluer.

Drake a également mis en évidence quelques mécanismes chimiques influençant le taux de mutation chez les virus à ARN, tels que l'oxydation ou la méthylation des nucléotides (Drake 1993), tandis que d'autre mettaient en évidence la désamination de nucléotides non appariés aboutissant à des mutations transitionnelles, c'est-à-dire à une substitution entre une cytosine et une thymine (Walsh 2006) ou entre une guanine et une adénine (Caride 2002). Cette désamination constitue un mécanisme dit d'hypermutations, vraisemblablement mis en place par les cellules hôtes pour lutter contre l'infection virale, car conduisant le plus souvent à des mutations létales pour le virus. Ces mécanismes de défenses des organismes parasités pourraient d'ailleurs bien être à l'origine d'un biais dans la composition en nucléotides des génomes des parasites, souvent plus riches en AT qu'en GC. Notons cependant que la désamination des nucléotides peut également être un phénomène chimique spontané agissant sur une molécule d'ADN lorsqu'elle n'est pas appariée à son brin homologue, ce qui pourrait être à l'origine du taux de mutation élevé observé chez les virus à ADN simple brin (Drake 1991). Une autre hypothèse du taux de mutation élevé détecté chez ces virus pourrait être le fait que leur ADN n'étant pas méthylé, il ne puisse pas être pris en charge par le complexe enzymatique de réparation des erreurs de la cellule hôte (Arguello-Astorga 2007). Enfin, les structures secondaires du génome viral, telles que les structures en épingles à cheveux (observées par exemple chez le virus PPA dans les parties 5' et 3' terminales de son génome) pourraient conduire les complexes enzymatiques responsables de la réplication des acides nucléiques cellulaires à introduire des délétions, car n'étant pas à même de les décrypter (Pathak 1992).

La plupart des études réalisées jusqu'ici montraient une tendance des virus à ARN à évoluer plus rapidement que les virus à ADN. La réalité semblerait être cependant nettement moins tranchée, que cela soit le fait de l'architecture génomique des virus, de leur cycle réplicatif, de l'unicité ou la pluralité des hôtes qu'ils infectent, et qui les soumet à une sélection purificatrice plus ou moins drastique (Weaver & Reisen 2009), ou même de leur propre capacité à être infectés par d'autres virus (La Scola *et al.* 2008). Toutes ces données, loin d'être secondaires, affectent en réalité de façon critique la dynamique évolutive des virus.

Chez la plupart des virus à ARN étudiés, le taux de substitution, c'est-à-dire de mutations fixées dans la descendance, a été estimé de l'ordre 10^{-2} à 10^{-5} substitutions par site et par an (Jenkins *et al.* 2002), avec une moyenne de 10^{-3}, ce qui correspond à une dizaine de substitutions par site et par an, pour un virus dont la taille du génome d'ARN simple brin serait de 10 kb. Cela semble peu en termes de substitutions, quand bien même elles seraient non synonymes. Chez les virus à ARN entrainant une infection chronique de leur(s) hôte(s), l'évolution apparait comme beaucoup plus élevée au niveau intra-hôte qu'au niveau inter-hôtes. C'est vraisemblablement dû au fait que la survie du virus à l'intérieur d'un même hôte dépend grandement de sa capacité à échapper au système de défense, essentiellement représenté par le système immunitaire de son hôte. Le virus est donc soumis à une sélection positive (Nielsen & Yang 1998). De la même façon, un grand nombre de mutants sont éliminés lors de la transmission inter-hôtes, par le mécanisme de sélection purificatrice, engendré par la réponse immunitaire médiée par les lymphocytes T cytotoxiques, comme cela a été démontré en ce qui concerne les virus HIV et SIV (Li *et al.* 2007). En effet, ces virus, pour survivre dans leur hôte, doivent progressivement s'adapter à son système immunitaire, pour principalement échapper à la réponse des lymphocytes T cytotoxiques. Pour ce faire, les mutations sont principalement localisées sur les parties du génome dont les protéines codées sont associées avec le complexe HLA de l'hôte. Ainsi, lors d'une transmission inter-hôte, c'est-à-dire entre hôtes ayant des complexes HLA différents, la plupart des variants ne représentent plus un bénéfice pour le maintien de la population virale, et seront donc éliminé (Friedrich *et al.* 2004). Au final, trois facteurs sont déterminants pour expliquer la fréquence d'apparition et de fixation des mutants dans les populations de virus ARN à évolution rapide : le taux de mutation apparaissant spontanément au cours de la réplication des acides nucléiques, le temps de génération, et les relations intra et inter-hôtes du virus, au cours desquelles les mutations avantageuses seront fixées en de réelles substitutions. Un quatrième facteur semble aussi jouer un rôle déterminant : il s'agit de la taille effective de la population. Rappelons qu'au sens de la génétique des populations de Fisher et de Wright (Fisher 1930 ; Wright 1931), une population doit avoir une taille suffisante pour que la diversité qu'elle génère soit représentative de ce que l'on observerait si l'on étudiait une population beaucoup plus grande. Or, il semble que la rapidité avec laquelle les mutations avantageuses sont fixées dans la population soit proportionnelle avec la taille initiale de cette population. Cela est probablement dû aux interactions synergiques entre quasi-espèces. Chez les poliovirus, il a été montré que la présence au sein d'une population de virus non pathogènes d'une fraction de virus pathogènes ne suffisait pas pour établir le phénotype virulent de la population entière. Ceci atteste d'un seuil de virulence en deçà duquel la population de virus non pathogènes protègerait l'hôte de la virulence d'une fraction d'entre elle (Lancaster & Pfeiffer 2011). La dynamique virale (i.e. l'apparition des mutants), au niveau intra-population et donc tout particulièrement au sein d'un même hôte, serait ainsi un facteur absolument déterminant du spectre de la virulence virale (Lancaster & Pfeiffer 2012). On notera à titre indicatif que ce mécanisme de seuil de virulence peut tout à fait être applicable à d'autres

microorganismes tels que les bactéries. Il a même été mis en évidence dans le cas des cellules cancéreuses et même dans celui des prions (Ojosnegros *et al.* 2011). C'est dire combien la détermination des taux de substitutions chez les virus peut être déterminante en termes de contrôle, d'émergence et de transmission des maladies virales.

Nous l'avons noté, la relation virus-hôte est fondamentale en terme d'évolution. L'on peut même penser qu'il y a bien souvent une longue coévolution de l'hôte et de son parasite, sans pour autant que cela conditionne complètement le taux d'évolution des parasites (McGeoch *et al.* 2000). Néanmoins, ce n'est pas la co-spéciation qui est le mode dominant d'évolution des virus à ARN, mais le passage de la barrière d'espèce (Holmes 2003). L'un des virus emblématique utilisé pour justifier que la coévolution d'un virus à ARN avec son hôte n'entraine pas un taux d'évolution élevé, mais au contraire aurait plutôt tendance à le réduire, est celui du SFV (*simian foamy virus*), un rétrovirus ubiquitaire des primates (Meiering & Linial 2001). Le taux d'évolution de ce virus est de $1,7 \times 10^{-8}$ substitutions par site et par an (Switzer *et al.* 2005). Ce taux d'évolution très faible pour ce type de virus serait en rapport avec la latence du virus chez son hôte, c'est à dire le faible nombre de copies virales produites et le fait que le virus soit intégré au génome de son hôte. Cette explication prévaut également pour un autre virus intégratif, le HTLV-II (*human T-cell lymphotropic virus* type II) dont le taux d'évolution est de 10^{-7} au sein d'un même individu (coévolution) tandis qu'il est nettement supérieur ($>10^{-4}$) dans le cas d'une transmission entre individus différents au cours de laquelle il se réplique à haut niveau (Salemi *et al.* 1999).

Dans les deux cas que nous venons de citer, le virus se réplique via une rétro-transcriptase dont on sait qu'elle est plus fidèle que les ARN-polymérases ARN-dépendantes. Il existe pourtant des virus à ARN utilisant des ARN-polymérases ARN-dépendantes qui évoluent très lentement tels que chez les humains, le flavivirus GBV-C (Suzuki *et al.* 1999). Dans ce cas, la coévolution est l'explication la plus plausible, surtout lorsqu'on constate pour ce même virus que le passage d'un hôte à un autre entraine une très nette augmentation du taux d'évolution, ne serait-ce que pour permettre au virus d'optimiser le détournement de la machinerie cellulaire en s'adaptant (*fitness*) aux nouveaux complexes enzymatiques ou à la préférence codon du nouvel hôte. Les virus ARN semblent aussi évoluer plus lentement chez les plantes (Garcia-Arenal *et al.* 2001). Dans ce cas, l'explication la plus couramment utilisé est celle du « goulot d'étranglement » (*population bottleneck*), mécanisme qui advient au sein de populations et au cours duquel une partie importante des individus meurt ou est empêchée de se reproduire (Nei 2005).

Il apparait donc évident que la capacité des virus ARN à diverger n'est pas seulement intrinsèque, mais dépend fondamentalement de son histoire, c'est-à-dire de ses relations avec son hôtes.

Qu'en est-il des virus à ADN ?

Les virus à ADN ont en général un taux de mutation largement inférieur aux virus à ARN et aux rétrovirus. Cependant, il existe, des différences selon la nature de l'ADN génomique de ces virus, c'est-à-dire selon qu'il soit à simple ou à double brins. En effet, les virus à ADN simple brin montre un taux d'évolution proche de celui des virus à ARN, cela pouvant être dû, comme nous l'avons dit plus haut, au fait que des désaminations spontanées apparaissent sur les brins d'ADN lorsqu'ils ne sont pas appariés. De plus, contrairement aux virus à ARN, les virus à ADN simple brin, en général de petite taille (~13 kb), montrent un taux d'évolution élevé tant au niveau intra-hôte, qu'inter-hôtes (Isnard *et al.* 1998). Leur taux d'évolution élevé ne pouvant être attribué aux polymérases qui les répliquent, outre les désaminations, d'autres phénomènes tels qu'une forte propension à recombiner, doivent intervenir.

L'évolution de plusieurs virus à ADN simple brin a été particulièrement étudiée. Le CPV-2 (*canine parvovirus* type 2), le FPV (*feline panleukopenia virus*) dont dérive le CPV-2, ou encore le PCV2 (*porcine circovirus* type 2) ont un taux d'évolution de 10^{-4} pour les deux premiers virus (Shackelton *et al.* 2005) et de 1,2 x10^{-3} pour le circovirus (Firth *et al.* 2009). Chez les plantes, le taux d'évolution du begomovirus TYLCV (*tomato yellow leaf curl virus*) a été estimé à 2,88 x 10^{-4} (Duffy & Holmes 2008). Ces taux d'évolution sont donc très proches de ceux des virus à ARN.

L'évolution des virus animaux à ADN double brins apparait différente de celle des virus à ADN simple brin. Ils sont supposés avoir co-évolués, ou co-divergés, avec leurs hôtes depuis des temps très reculés, on parle de millions d'années, à cause de la nature de leur association, et de la pression de sélection qu'ils exercent l'un sur l'autre. La coévolution se traduit alors par un changement chez le virus comme chez l'hôte, comme cela a été le cas lors de l'introduction du virus de la myxomatose dans une population de lapins naïfs, qui a vu la virulence du virus diminuer tandis que la capacité de résistance de l'hôte s'accroissait (Woolhouse *et al.* 2002). Chez les papillomavirus humains (de petits virus à ADN double brins), dont les origines dateraient des primates pré-humains, la non transmission entre les espèces suggère une longue et lente coévolution dans leur hôte avec un taux d'évolution estimé autour de 10^{-8} substitutions par site et par an (Bernard 1994). Pour les gamma-herpes virus infectant les vertébrés, le taux d'évolution a été déterminé autour de 10^{-9}, soit un taux très proche de celui de leurs hôtes (McGeoch & Gatherer 2005 ; McGeoch *et al.* 2005). La date de leur spéciation remonte quant à elle, à plus ou moins 200 millions d'années (McGeoch *et al.* 1995). Ces données de spéciation des virus à ADN double brins ont été calculées selon l'hypothèse d'une coévolution avec leurs hôtes. Par exemple, le polyomavirus JCV est supposé avoir co-divergé lors de la migration des populations hors d'Afrique il y a environ 200 000 ans (Pavesi 2005), et sous l'hypothèse de coévolution, son taux d'évolution a été estimé à 10^{-7} substitutions par site et par an (Hatwell & Sharp 2000). Or, les phylogénies du virus et de ses hôtes ne semblent pas en concordance, l'échantillonnage sériel des souches virales produisant un taux d'évolution plus élevé. Il est alors difficile d'affirmer que les taux déterminés sont les vrais taux d'évolution de ces virus.

Les molécules d'ADN étant supposées être postérieures aux molécules d'ARN (Forterre 2005), les virus à ADN doivent logiquement être postérieurs aux virus à ARN. De plus, le monde du vivant qui est parvenu jusqu'à nous étant basé sur la réplication de l'ADN pour assurer sa continuité, les virus à ADN pourraient bien avoir des origines communes. Un groupe de virus à ADN double brins a été particulièrement étudiée : les grands virus à ADN nucléocytoplasmiques (NCLDV, *nucleocytoplasmic large DNA viruses*). Ces virus ne se répliquent que dans le cytoplasme des cellules qu'ils infectent. Même si parfois le cycle de leur réplication commence dans le noyau, l'achèvement de leur morphogénèse est toujours cytoplasmique (Iyer *et al.* 2006). De plus, ils sont relativement indépendant des cellules qu'ils infectent, car leur grand génome (de 100 kb à 1.2 Mb) code pour un nombre important de gènes, dont de nombreuses enzymes nécessaires à leur propre multiplication : polymérases et hélicases pour la réplication, topoisomérases pour la manipulation de l'ADN, facteurs de l'initiation et de l'élongation de la transcription de l'ARN, ATPases pour le conditionnement de l'ADN, ainsi que des protéines chaperonnes pour la conformation des protéines servant à l'encapsidation (Iyer *et al.* 2001). Bien qu'ils partagent certains gènes typiquement viraux avec d'autres grands virus à ADN comme les herpesvirus et les baculovirus (Koonin *et al.* 2006), ils ont un noyau commun de gènes qui les distinguent clairement des autres classes de virus (Iyer *et al.* 2006). Ainsi, si la phylogénie de cette famille et celles de ses principaux membres ont été très étudiées, c'est que cette famille est supposée être monophylétique, c'est-à-dire qu'elle dériverait d'un même ancêtre commun (Iyer *et al.* 2001). Bien des phylogénies ont donc été réalisées pour démontrer les parentés entre les six familles de virus appartenant aux NCLDV (Iyer *et al.* 2006 ; Koonin & Yutin 2010 ; Yutin & Koonin 2009 ; Yutin *et al.* 2009). Ces familles sont les *Poxviridae*, les *Iridoviridae* et les *Ascoviridae*, les *Phycodnaviridae*, les *Mimiviridae* et les *Asfarviridae*. Les *Marseilleviridae* récemment découverts (Boyer *et al.* 2009) pourraient bien constituer une septième famille parmi les NCLDV (Koonin & Yutin 2010). Malgré leur monophylie, ces virus infectent un large spectre d'hôtes : amibes, algues, coraux, invertébrés et insectes, reptiles, vertébrés mammifères ou non mammifères. Ils infectent l'ensemble du spectre du règne eucaryote, suggérant une coévolution ne remontant pas uniquement à leur(s) hôte(s) respectif(s) actuel(s), mais possiblement à l'ancêtre commun du règne eucaryote, ce qui suppose une très longue coévolution.

Parmi les six (ou sept) familles composant les NCLDV, une seule possède un membre unique : la famille des *Asfarviridae*. Ce membre est du genre *Asfivirus*, c'est le virus ASFV (*African swine fever virus*), responsable de la Peste porcine africaine (PPA). C'est ce virus qui nous a servi de modèle d'étude dans ce travail de thèse.

6- La Peste porcine africaine (PPA), ou African swine fever (ASF)

6-1- Historique – Distribution géographique

Le berceau de la Peste porcine africaine est l'Afrique de l'Est, dans la région des grands lacs. Elle fut décrite pour la première fois par en 1921 (Montgomery 1921), suite à deux premières épidémies survenues au Kenya en 1903 et en 1906, qui ont été suivies d'une quinzaine de foyers épidémiques entre 1909 et 1915. Ces épidémies ont alors tué jusqu'à 99% des porcs domestiques atteints. Elle a par la suite été détectée en Afrique du Sud puis en Angola, tout en restant confinée sur le continent africain. Elle a quitté l'Afrique pour la première fois en 1957, partant de l'Angola pour atteindre l'Europe, par le Portugal, où la maladie a entrainé la mort de 100% des animaux atteints. Après 3 ans d'accalmie, la maladie réapparait au Portugal en 1960, d'où elle gagne l'Espagne. Par suite, on la détectera en France (1964), en Italie (première apparition en 1967, puis réémergences en 1969 et en 1973), à l'île de Malte (1978), en Belgique (1985) et aux Pays-Bas (1986). Au Portugal, elle restera présente de 1960 à 1993 puis réémergera une dernière fois en 1999. Depuis lors, elle a été éradiquée de tous les pays européens excepté en Sardaigne, où elle est devenue endémique depuis son entrée en 1978 (Arias & Sánchez-Vizcaíno 2002b). Depuis l'Europe, elle a traversé l'océan Atlantique pour être introduite pour la première fois à Cuba en 1971, puis au Brésil en 1978, d'où elle atteint les Caraïbes : République Dominicaine en 1978, Haïti en 1979, et à nouveau Cuba, en 1980. Afin qu'elle ne puisse se répandre au-delà des foyers épidémiques, l'éradication a été réalisée, dans ces zones par un abattage systématique des cheptels porcins (Arias & Sánchez-Vizcaíno 2002a).

En Afrique, on pensait que la zone où sévissait la PPA était restreinte à l'Afrique Centrale, de l'Est et du Sud. Or, il s'avère que le Cap Vert a probablement été atteint dès les années 1960, puis le Nigéria en 1973, le Sénégal depuis au moins 1978 (pays dans lequel elle est aujourd'hui considérée comme endémique) et le Cameroun, en 1982. Depuis lors, la maladie est devenue une réelle menace pour toute l'Afrique de l'Ouest, avec des foyers épidémiques déclarés en Côte d'Ivoire (1996), au Benin, au Togo et au Nigeria (1997) ou encore au Ghana, en 1999. Elle est revenue au Kenya et au Mozambique en 1994 (FAO 2002), pour enfin atteindre Madagascar pour la première fois en 1998 (Gonzague *et al.* 2001) et l'île Maurice en 2007 (Lubisi *et al.* 2009). La PPA s'est donc propagée à quasiment toute l'Afrique sub-saharienne.

Depuis 2007, la PPA a refait son apparition sur le continent Européen par la Géorgie (Rowlands *et al.* 2008), s'est répandue dans le Caucase (Arménie, Azerbaïdjan) puis a atteint la Fédération de Russie où elle s'est maintenant propagée. Le virus responsable de ces épidémies a d'ailleurs été caractérisé comme appartenant au même groupe que le virus malgache (Malogolovkin *et al.* 2012). La figure 7 montre les pays endémiques, épidémiques, et nouvellement atteints.

La PPA est une maladie infectieuse, très contagieuse et spécifique des suidés. De par son très grand pouvoir de diffusion, elle fait partie des maladies à déclaration obligatoire par l'Organisation Mondiale de la Santé Animale (Office International des Epizooties, OIE). Elle est asymptomatique chez les suidés sauvages africains tels que les phacochères, les potamochères ou les hylochères, qui peuvent donc être considérés comme les réservoirs de la maladie. De plus, c'est une maladie vectorielle puisqu'elle peut être transmise aux animaux par un arthropode argaside, une tique molle du genre *Ornithodoros* (Plowright 1977).

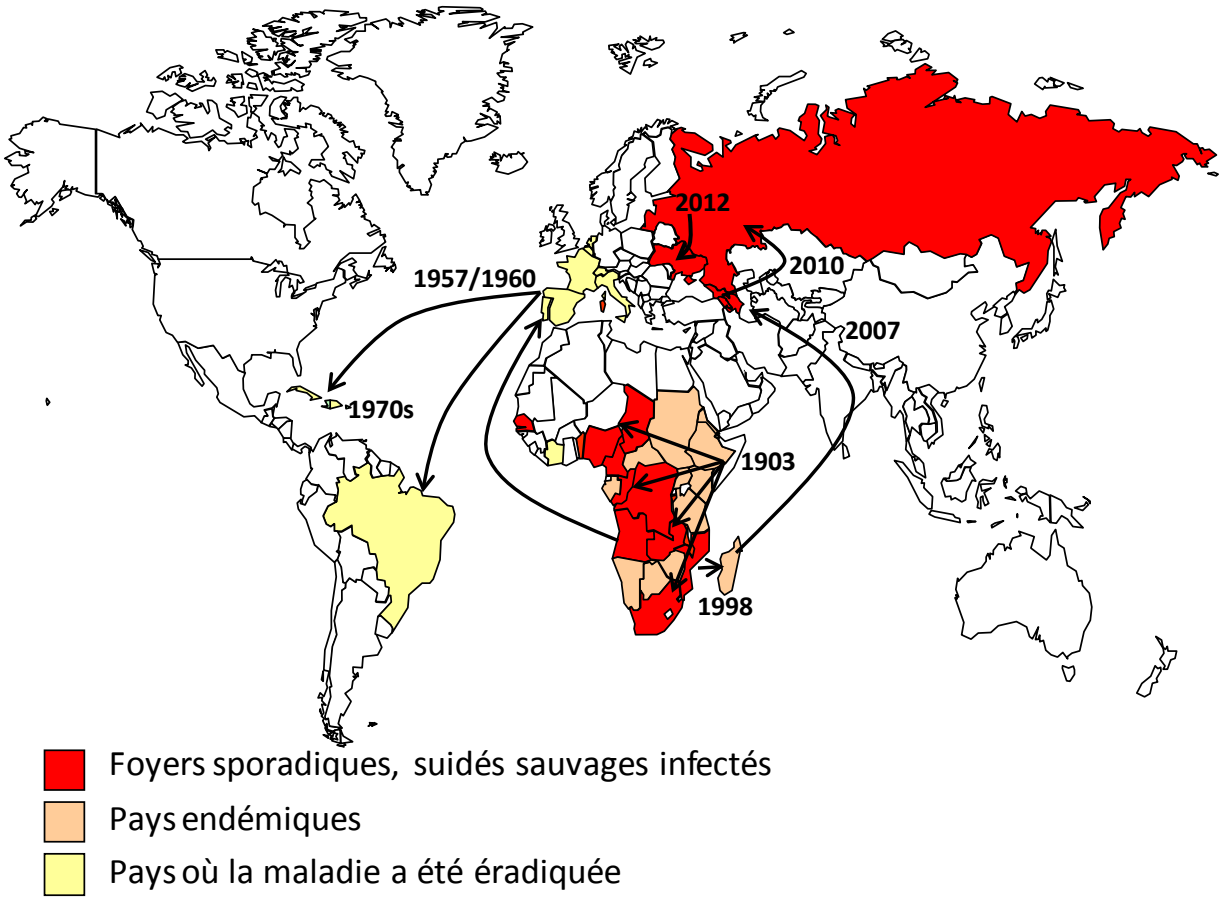


Figure 7 : Répartition de la Peste porcine africaine dans le monde. De 1903 à 1957 la maladie est restée confinée au continent africain. En 1957, partie d'Angola, elle a atteint l'Europe par le Portugal, s'est répandue dans plusieurs pays européens, puis a gagné l'Amérique du Sud et les Caraïbes au cours des années 1970. Elle a été éradiquée dans tous les pays atteints sauf en Sardaigne, où elle est devenue endémique. Depuis 2007, la PPA est revenue en Europe par le Caucase, d'où elle a atteint la fédération de Russie (2010) puis l'Ukraine (2012).

6-2- Signes cliniques – Pathogénie

Selon les isolats viraux responsables des foyers épidémiques, une large gamme de symptômes a été observée chez les porcs infectés : infection suraigüe, aigüe, subaigüe et portage chronique (Moulton & Coggins 1968). Les formes suraigües et aigües provoquent la mort de près de 100% des animaux infectés, en général dans les 13 jours suivant l'infection. Cependant, les individus survivants deviennent porteurs du virus sur plusieurs mois voire années, devenant ainsi de vraies « bombes biologiques ». Dans sa forme subaigüe, la maladie dure entre 15 et 45 jours, provoquant fièvres récurrentes, toux, dyspnée, anorexie et avortements spontanés des femelles gestantes. Le taux de mortalité induit varie entre 30 et 70%. Enfin, la forme chronique provoque moins de 30% de mortalité et dure de 2 à 5 mois.

La physiopathologie de la maladie dépend de la forme de l'infection. Dans le cas de la forme suraigüe, les animaux meurent généralement avant l'apparition de signes cliniques. Néanmoins, des rougeurs peuvent apparaitre sur la peau de l'animal, ainsi qu'une congestion et une fibrination des organes (Penrith *et al.* 2004). Les formes aigües de la maladie produisent de nombreuses lésions, parmi lesquels les plus marquants sont : cyanose des extrémités et du ventre de l'animal, avec des ecchymoses ; hémorragies et congestion des muqueuses avec exsudations oculaires et écoulements naseaux. L'analyse *post-mortem* montre une hémorragie généralisée des organes avec apparitions de pétéchies sur les muqueuses, une splénomégalie et des œdèmes pulmonaires ainsi qu'une hyperplasie du système lymphatique (Hess 1981 ; Penrith *et al.* 2004). Les formes subaigües, quant à elles, sont caractérisées par des œdèmes articulaires, une péricardite effusive ainsi qu'une inflammation des ganglions lymphatiques. A noter que les animaux développant cette forme de la maladie sont souvent sujets aux pneumonies. Enfin, dans les formes chroniques, on note des éruptions cutanées pouvant conduire à une nécrose épidermique ainsi que de nombreux abcès purulents au niveau de la tête et des tétines. Se développent également : une arthrite purulente au niveau des articulations, une hépatisation nécrotique des poumons, une péricardite, une splénomégalie et une inflammation des ganglions lymphatiques (Penrith *et al.* 2004).

6-3- Prévention de la maladie

La PPA est la maladie la plus grave touchant le porc domestique, catastrophe sanitaire pour les élevages porcins, principalement dans les pays du Sud pour lesquels le porc est une source peu coûteuse de production de protéines animales et dont la demande est en constante progression. C'est également désormais une menace pour les élevages des pays caucasiens et de la Fédération de Russie, c'est-à-dire une menace potentielle pour l'Europe. Sans vaccin ni thérapie disponibles, le contrôle de la maladie repose sur un diagnostic

précoce et sur des mesures sanitaires de biosécurité : abattage et incinération des cheptels atteints et/ou exposés, ainsi qu'un contrôle drastique des échanges commerciaux avec les pays concernés. En effet, le commerce des produits à base de porcs est strictement interdit, car potentiellement contaminants. Néanmoins, le diagnostic différentiel est parfois difficile, car la PPA partage de nombreux symptômes avec d'autres maladies hémorragiques telles que la peste porcine classique. Enfin, l'augmentation exponentielle des mouvements d'animaux et du commerce de produits carnés accroît les risques d'introduction de la maladie dans les pays indemnes.

Toutefois, l'éradication a déjà été possible dans des zones où la maladie était devenue endémique, comme la péninsule ibérique, les Caraïbes ou le Brésil. De nombreux pays sont cependant toujours en proie à la maladie. La principale difficulté pour son contrôle, puis son éradication, est avant tout l'existence de porcs en divagation, pouvant s'infecter par contact direct avec les suidés sauvages réservoirs de la maladie ou avec le vecteur tique, qui peut rester infectant durant de longues années (Greig 1972 ; Plowright *et al.* 1970b). Qui plus est, l'endémisation de la maladie entraine l'apparition de porteurs chroniques asymptomatiques, qui ne seront donc pas détectés lors de l'abattage des animaux (Sanchez-Vizcaino 2006), entrainant ainsi un risque de propagation.

Jusqu'ici, les stratégies vaccinales qui ont été mises en place, qu'elles soient basées sur l'utilisation de souches atténuées, inactivées, ou de protéines recombinantes se sont avérées inefficaces (Barderas *et al.* 2001). Cela résulte vraisemblablement de la grande variabilité antigénique du virus et de l'absence d'induction d'une réponse neutralisante efficace par le système immunitaire des porcs après vaccination. Néanmoins, l'utilisation de souches atténuées comme la souche ASF/NH/P68 a permis l'immunisation d'animaux contre une épreuve infectieuse avec des virus homologues (Leitao *et al.* 2001). La vaccination n'a en revanche jamais protégé les animaux contre une épreuve infectieuse avec des souches hétérologues (Ruiz Gonzalvo *et al.* 1986b), sauf lorsque cette épreuve intervenait après un essai de protection homologue (King *et al*., 2011).

Les coûts socio-économiques engendrés par la PPA, en particulier dans des pays déjà souvent économiquement en difficulté, sont tels que la préservation des zones encore indemnes est absolument capitale.

7- Le virus de la Peste porcine africaine

7-1- Taxonomie – Classification

La PPA est due à un virus à ADN double brin enveloppé, de symétrie icosaédrique, le virus ASFV (African *swine fever virus*), ou virus PPA. Sa réplication, quasi exclusivement cytoplasmique le place d'emblée parmi les grands virus à ADN nucléocytoplasmiques

(NCLDV) (Iyer *et al.* 2006). Ses caractéristiques morphologiques ont été responsables de sa classification initiale parmi les *Iridoviridae*. Cependant, l'organisation de son génome, telles que les structures en « épingles à cheveux » (boucles d'ADN formées de séquences répétées inversées et liées de façon covalentes) situées à chacune de ses extrémités (De la Vega *et al.* 1994) ou encore la présence de gènes codant pour la machinerie nécessaire à une partie de la réplication de son ADN ou à la maturation des ARN messagers précoces, le rendait plus proche des *Poxviridae*. Néanmoins, des analyses phylogénétiques étudiant les NCLDV ont discriminé le virus PPA des *Iridoviridae* et des *Poxviridae* (Raoult *et al.* 2004). Au final, des caractéristiques architecturales ainsi qu'une information génomique propres ont amené le Comité International de Taxonomie des Virus (ICTV) à créer une nouvelle famille pour le classer : la famille des *Asfarviridae*, dans laquelle il est membre unique, dans le genre *Asfivirus* (Dixon *et al.* 2005). De plus, il est important de noter que le virus PPA est le seul arbovirus (ou *arthropod-born virus*) à ADN décrit jusque-là car il infecte les tiques molles du genre *Ornithodoros* (Plowright 1977 ; Wardley *et al.* 1983).

7-2- Structure – Génome – Protéines codées

Les particules virales ont la forme d'un icosaèdre dont la structure est composée de plusieurs couches successives. Au centre du virus se trouve un cœur de 80 nm contenant un nucléoide de 30 nm contenant le génome viral. Ce cœur est entouré d'une première couche lipidique (enveloppe interne), surmontée d'une couche protéique formant une capside icosaédrique de 170 à 190 nm. La capside est faite de l'assemblage de 1892 à 2172 capsomères hexagonaux de 13 nm de long. Enfin, les virions extra cellulaires acquièrent une membrane externe au cours du bourgeonnement au travers de la membrane plasmique. La taille finale du virus est donc comprise entre 175 et 210 nm (Carrascosa *et al.* 1984 ; Rouiller *et al.* 1998). Les virions extra cellulaires matures contiennent les enzymes nécessaires à l'expression des gènes précoces, se faisant dès la pénétration du virus dans la cellule.

L'analyse des génomes complets d'isolats a permis de déterminer au moins 150 gènes potentiels codés par le génome viral. Situés sur les deux brins de l'ADN, ces gènes peuvent être lus dans les deux sens (Dixon *et al.* 1994 ; Yanez *et al.* 1995). De plus, certains gènes, comme le gène KP177R (codant pour la protéine de surface p22), peuvent être présents sous forme de copies multiples dans le génome (Chapman *et al.* 2008). Les analyses biochimiques des particules virales ont montré la présence d'au moins 54 protéines de structures (Carrascosa *et al.* 1985). Certaines comme les protéines p150, p37, p34 et p14 (produits du clivage de la polyprotéine pp220) ont été localisées dans le cœur même du virion. De même, la protéine VP72 (codée par le gène B646L) est localisée à la surface des virions intracellulaires et constitue la protéine majeure de la capside virale qui entoure la membrane interne des virions non enveloppés. Elle n'est cependant pas une véritable protéine membranaire (Cobbold *et al.* 2001). L'enveloppe externe des virus extracellulaires

contient les protéines d'attachement p12 et p24 (Alcami *et al.* 1992) et la protéine CD2v, seule protéine glycosylée de la particule virale, qui est responsable de l'hémadsorption des globules rouges (Ruiz-Gonzalvo *et al.* 1996). Elle contient également la protéine p32 (encodée par le gène CP204L), une phosphoprotéine sous forme hexamèrique lorsqu'elle est intégrée à la membrane, ainsi que la protéine p54 (codées par le gène E183L), la plus importante des protéines constitutives de la membrane et localisée à la surface externe des virus extracellulaires. L'enveloppe externe des virus extracellulaires est très complexe puisque, sur sa face interne elle contient également des protéines possédant des domaines transmembranaires, telles que les protéines J18L, p12 ou p17 (Alcami *et al.* 1992 ; Sun *et al.* 1996). Enfin, les particules virales contiennent plusieurs types d'enzymes participant à la réplication du génome viral, une ARN-polymérase ADN-dépendante ayant un rôle dans l'initiation et l'élongation de la transcription, ainsi que la maturation des ARN messagers. On trouve également dans le virus une kinase, une nucléoside phosphohydrolase, une phosphatase acide ainsi que deux désoxyribonucléases ayant une action sur l'ADN simple brin (Yanez *et al.* 1995) (Figure 8).

A l'intérieur du nucléoide, se trouve le génome viral. Il est constitué d'un double brin d'ADN riche en A-T (35 à 38% de GC selon les différents isolats séquencés) d'une longueur comprise entre 170 et 193 kb selon les isolats considérés (Chapman *et al.* 2008). Ces différences de taille proviennent de l'insertion (ou de la délétion), lors de la réplication du génome, d'indels de séquences répétées en tandem. Ces séquences codent pour cinq familles mutigéniques (MFG pour *multi-genes family*) : MGF 360, MGF 110, MGF 300, MGF 530 (ou 505) et MGF 100 (de la Vega *et al.* 1990 ; Gonzalez *et al.* 1990 ; Yozawa *et al.* 1994). La structure des extrémités du génome est en épingle à cheveux, c'est-à-dire que les extrémités des brins d'ADN sont liées de façon covalente par 37 nucléotides, majoritairement composés de T et de A, montrant un appariement incomplet (Gonzalez *et al.* 1986). Dans le prolongement direct de ces structures terminales en épingle à cheveux, on trouve des séquences répétées inversées sur une longueur de 2,1 à 2.5 kb (Yanez *et al.* 1995). La partie centrale du génome est constituée de répétitions inversée en tandem de 33 à 34 nucléotides, motifs présents également dans d'autres parties du génome (Dixon *et al.* 1994). Enfin, il existe une partie hypervariable d'environ 400 nucléotides, située au sein de la partie centrale très conservée de 125 kb du génome (Sumption *et al.* 1990).

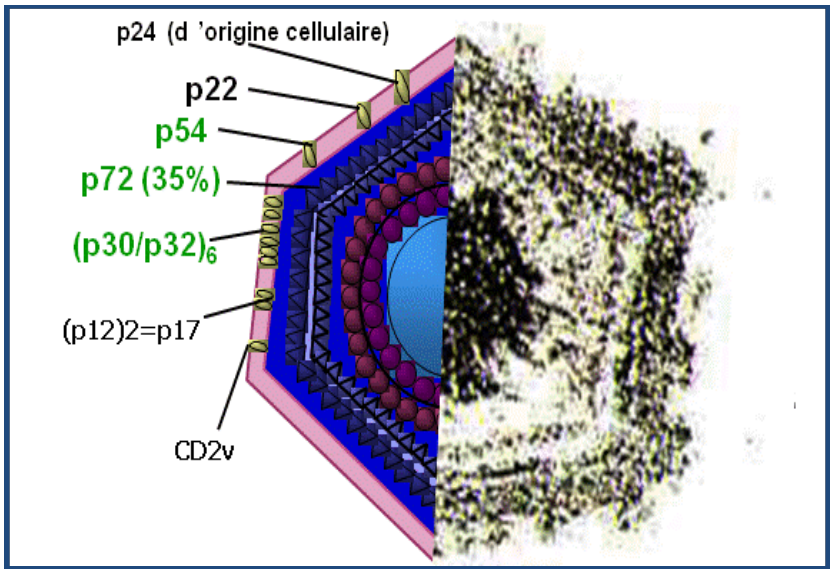


Figure 8 : Structure du virus de la Peste porcine africaine. Les protéines majeures sont indiquées : VP72, la protéine majeure (MCP) d'une capside de 80 nm, entourant le nucléoïde qui contient le génome viral. Avec les protéines membranaires p54 et p32 ce sont les plus immunogènes des protéines virales. La forme du virus est icosaédrique. Les virions acquièrent une enveloppe externe lors du bourgeonnement à la membrane cellulaire pour atteindre une taille de 175 à 215 nm.

7-3- Pénétration dans la cellule – Réplication – Morphogénèse

Le virus PPA infecte préférentiellement les monocytes et les macrophages. On note cependant que la sensibilité de ces cellules au virus dépend de leur stade de maturation (Minguez *et al.* 1988 ; Wardley *et al.* 1977). Selon le stade de l'infection, on trouvera également le virus dans les cellules dendritiques et endothéliales (Vallee *et al.* 2001), les mégacaryocytes, les neutrophiles, les hépatocytes et même les plaquettes (Carrasco *et al.* 1992 ; Fernandez *et al.* 1992a ; Fernandez *et al.* 1992b).

La pénétration du virus dépend en premier lieu de sa fixation sur les cellules, via un ou des récepteurs membranaires. Jusqu'ici, toutes les cellules permissives au virus PPA expriment les récepteurs SWC9 et CD163 (un marqueur de maturation des monocytes, (McCullough *et al.* 1999 ; Sanchez-Torres *et al.* 2003 ; Sanchez *et al.* 1999). Une seule protéine d'attachement du virus a été identifiée, il s'agit de la protéine d'enveloppe p12. Cette protéine, synthétisée tardivement lors de l'infection et montrant un haut degré de conservation entre les isolats (Angulo *et al.* 1992), contient un domaine transmembranaire hydrophobe qui s'ancre au sein de l'enveloppe virale (Angulo *et al.* 1993). Une fois fixé, le virus pénètre dans la cellule par endocytose via les endosomes puis les lysosomes (Valdeira & Geraldes 1985) (Figure 9).

La réplication du virus PPA commence dès son entrée dans la cellule par la transcription de gènes très précoces, possiblement des facteurs de l'initiation de la

transcription (Salas 1999), grâce à l'ARN-polymérases-ADN-dépendante présente dans le virus (Yanez *et al.* 1993a). Ces facteurs essentiels à la transcription précoce des ARNm incluent des protéines homologues à celles codées par les poxvirus, telles que les protéines A2L (B385R pour le virus PPA), A7L (G1340L), qui n'ont pas d'équivalent dans le règne eucaryote ni même parmi les autres virus, et sont requises pour la reconnaissance des promoteurs viraux (Iyer *et al.* 2006). Les ARNm sont ensuite polyadénylés et coiffés par des enzymes également présentes dans le virion (Yanez *et al.* 1993a). Des ARNm précoces et intermédiaires vont être également transcrits mais rester silencieux jusqu'au début de la réplication proprement dite. Enfin, des ARNm tardifs seront synthétisés, suggérant une régulation en cascade de l'expression des gènes. De plus, la structure terminale des ARNm précoces est constituée de sept thymines, au lieu des dix habituellement requis comme signal de fin de traduction, suggérant la possibilité d'ARNm polycystroniques. Longs d'une cinquantaine de nucléotides AT riches, les promoteurs précoces sont placés en amont du codon d'initiation (Almazan *et al.* 1992 ; Almazan *et al.* 1993 ; Goatley *et al.* 2002).

Comme pour tous les virus NCLDV, la réplication du virus a lieu dans le cytoplasme cellulaire, dans des endroits périnucléaires discrets appelés « usines à virus » (Garcia-Beato *et al.* 1992 ; Rojo *et al.* 1999). Cette synthèse nécessite tout d'abord la présence d'enzymes virales de réplication telles que polymérase virale (vraisemblablement codée par le gène G1207R) et protéine homologue aux ADN polymérases α-like (Rodriguez *et al.* 1993b). Elle nécessite également des enzymes de manipulation des acides nucléiques : une thymidine kinase, une ribonucléotide réductase ainsi qu'une ligase (Cunha & Costa 1992 ; Martin Hernandez & Tabares 1991 ; Yanez *et al.* 1993b ; Yanez & Vinuela 1993). Enfin, des enzymes de maturation et de réparation de l'ADN sont codées par le virus : une endonucléase apurinique/apyrimidique (APE), une polymérase de type β (Pol X) (Oliveros *et al.* 1999 ; Oliveros *et al.* 1997), ainsi qu'une ligase ATP-dépendante. Ces trois enzymes forment un complexe permettant l'excision et la réparation des bases (*base excision repair*, BER) (Sobol *et al.* 1996) et sont caractéristiques du virus ASFV. L'APE est une protéine homologue aux endonucléases de type IV (Lamarche & Tsai 2006), dont l'activité permet de catalyser l'hydrolyse en 5' des résidus ribophosphates de fragments d'ADN dénaturés. Ainsi génère-t-elle des fragments d'ADN dans lesquels ne manquent qu'un seul nucléotide, et dont les extrémités 3'-hydroxyle et 5'-2-désoxyribose-5'-phosphate sont directement utilisables par une polymérase. Il a toutefois été démontré que la polymérase X virale était structurellement et fonctionnellement unique en son genre (Showalter *et al.* 2001 ; Showalter & Tsai 2001). Avec seulement 174 acides aminés et l'absence de domaine N-terminal responsable de la liaison avec l'ADN (en comparaison avec les polymérases β homologues), cette enzyme est la plus courte des polymérases décrites à ce jour. De plus, lors de son activité de réparation de l'ADN (consistant à combler le manque d'un nucléotide dans l'ADN néosynthétisé par rapport au brin matrice), la Pol X induit de nombreuses erreurs de réplication, particulièrement la formation de mésappariements G-G, et ce avec la même efficacité que l'appariement canonique G-C (Showalter & Tsai 2001). Cette inclination à intégrer des erreurs au cours de la réplication du génome viral donne à cette enzyme un rôle

potentiellement mutagène. De plus, la présence de ces erreurs requiert que la ligase chargée d'assembler les fragments d'ADN puisse lier efficacement ensemble des extrémités contenant des mésappariements, c'est-à-dire une ligase elle aussi peu fidèle (Lamarche *et al.* 2005). Sans cette infidélité de la ligase virale, les erreurs provoquées par la Pol X pourraient s'avérer létales pour le virus. Ces deux enzymes de réparation infidèles pourraient donc être à l'origine d'une part de la variabilité du virus PPA. Un tel système génèrerait en effet un taux de mutation par génome (μ_g) compris entre 1,9 x 10^{-4} (Garcia-Escudero *et al*. 2003) et 1,1 x 10^{-1} (Lamarche *et al*. 2006), soit de 33 à 20 000 substitutions par génome répliqué. Avec des titres viraux pouvant atteindre 10^8 TCID50 chez le porc domestique, le nombre de variants ainsi générés serait considérable, si le rôle de la Pol X n'était pas restreint à la réparation de l'ADN, c'est-à-dire à la synthèse de petits fragments d'ADN. Le virus PPA montrant une variabilité assez élevée, son rôle dans l'émergence de mutants ne peut pas être négligé.

Même si la réplication a lieu quasi exclusivement dans le cytoplasme, le noyau cellulaire semble cependant jouer un rôle important dans la synthèse de l'ADN viral (Garcia-Beato *et al.* 1992). On trouve en effet dès les premiers stades de l'infection, des fragments courts d'ADN viral dans le noyau cellulaire, fragments qui pourraient jouer le rôle de précurseurs pour l'assemblage du génome complet dans le cytoplasme (Rojo *et al.* 1999). Le transport de ces molécules au travers de la membrane nucléaire reste mal connu, bien que la protéine p32, codée par le gène CP204L semble avoir un rôle dans cette translocation, via son interaction avec la ribonucléoprotéine cellulaire K au cours de l'infection (Hernaez *et al.* 2008).

Quatre-vingt cinq protéines virales ont été détectées lors de l'infection (Carrascosa *et al.* 1985 ; Carrascosa *et al.* 1984), dont deux, les polyprotéines pp62 et pp220 sont clivées au niveau d'un site d'acides aminés Gly-Gly-X (Lopez-Otin *et al.* 1989) pour donner naissance à pas moins de 6 protéines de structure : les protéines p35 et p15 et les protéines p150, p37, p34 et p14 (Andres *et al.* 1993 ; Simon-Mateo *et al.* 1997 ; Simon-Mateo *et al.* 1993). Certaines de ces protéines seront myristoylées tandis que d'autres seront phosphorylées (Afonso *et al.* 1992).

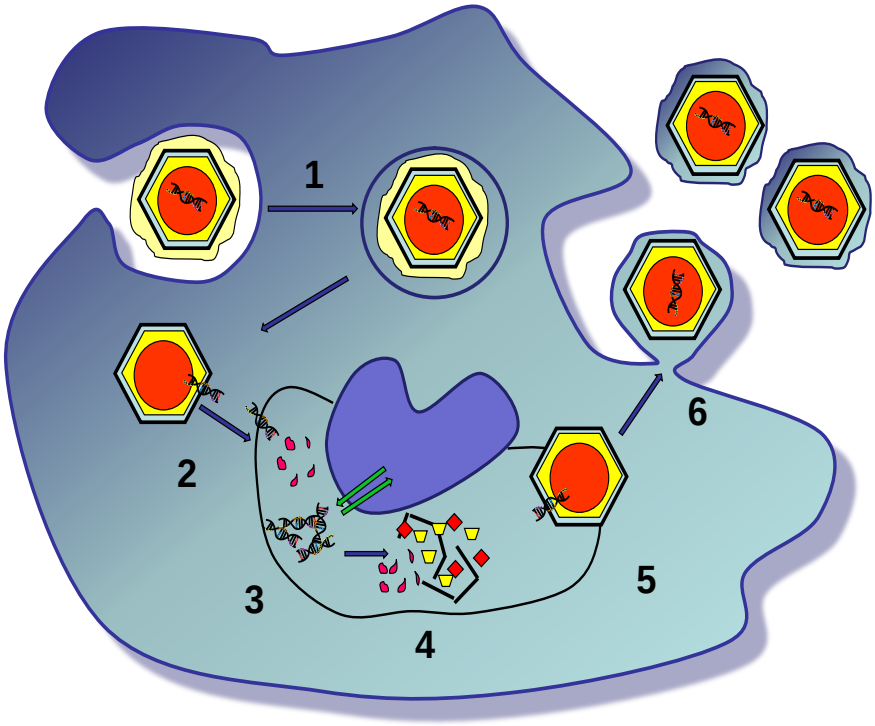


Figure 9 : Cycle réplicatif du virus PPA. (1) : pénétration du virus dans la cellule via les lysosomes après accrochage par le(s) récepteur(s). (2) : libération de l'ADN viral ainsi que des enzymes contenues dans le virion et permettant la traduction des gènes viraux précoces. La traduction des gènes précoces est indépendante des ARN-polymérases cellulaires. Les traduits précoces sont essentiels à la suite de la réplication du virus. (3) : réplication cytoplasmique du génome viral dans des « usines à virus » périnucléaires. La taille du génome varie entre 170 et 190 kb. (4) : expression des gènes intermédiaires et tardifs selon l'avancement de la réplication du génome. Les gènes tardifs codent pour des protéines de structure et des enzymes qui seront empaquetées dans le virion néo-formés. (5) : morphogénèse des virions. (6) : sortie des virions néo-formés par bourgeonnement à la membrane plasmique cellulaire. C'est au cours du bourgeonnement que les virions acquièrent leur enveloppe externe.

La morphogénèse du virus a donc lieu dans des usines à virus, sites périnucléaires proches du réseau de microtubules de l'appareil de Golgi et entourés de citernes de réticulum endoplasmique, de mitochondries et d'un réseau de vimentine (Andres *et al.* 1997 ; Heath *et al.* 2001 ; Rojo *et al.* 1998). On trouve à l'intérieur de ces sites toutes les formes intermédiaires de virus, de la structure membranaire incomplète au virus complet néoformé (Rouiller *et al.* 1998). Trois protéines sont fondamentales pour la morphogénèse du virus : les polyprotéines pp62 et pp220, dont le clivage produira six protéines structurelles (Andres *et al.* 2002a ; Andres *et al.* 2002b ; Andres *et al.* 1998 ; Andres *et al.* 1997) et la protéine VP72, composante majoritaire de la capside virale (Garcia-Escudero *et al.* 1998). A la fin de la morphogénèse, les particules virales néoformées sortent des cellules par bourgeonnement à la membrane plasmique (Arzuza *et al.* 1992), bourgeonnement au cours duquel elles acquièrent leur enveloppe externe. Les nouveaux virions peuvent également être libérés lors de l'apoptose et de la mort cellulaire induites par le virus. Ces derniers ne seront donc pas enveloppés.

7-4- Réponse immune – Virulence

Les porcs infectés, lorsqu'ils survivent à une infection par le virus PPA, développent une protection à long terme contre une nouvelle infection par un virus homologue, tandis qu'ils restent sensibles à une infection par un virus hétérologue (Wardley *et al.* 1985 ; Wardley & Wilkinson 1985). Alors qu'il avait été établi que l'infection par le virus PPA n'entrainait pas de production d'anticorps sériques protecteur chez l'animal (DeBoer *et al.* 1969 ; Ruiz Gonzalvo *et al.* 1986a), il a depuis été démontré que les anticorps pouvaient cependant jouer un rôle protecteur contre une réinfection (Zsak *et al.* 1993), y compris par transfert passif de sérum immun (Onisk *et al.* 1994).

De nombreuses protéines virales induisent des anticorps chez l'hôte, mais trois d'entre elles sont particulièrement immunogéniques : la protéine de capside VP72 et les protéines membranaires p54 et p32. Les anticorps neutralisants dirigés contre la VP72 et la p54 agissent durant la phase d'attachement du virus à la cellule, tandis que ceux dirigés contre la p32 inhibent l'internalisation du virus (Gomez-Puertas *et al.* 1996) sans cependant permettre une protection efficace (Neilan *et al.* 2004). Il a été toutefois démontré que l'utilisation d'un anticorps monoclonal reconnaissant un épitope de la protéine VP72 était capable de neutraliser, *in vitro*, quasiment complètement le virus (Borca *et al.* 1994). De la même manière, l'immunisation par un mélange de protéines p32 et p54, ou par une protéine chimérique p54/p32, a permis la neutralisation quasi complète du virus, avec une modification de la pathologie chez le porc vacciné, modification allant d'un retard dans les symptômes à une protection complète (Barderas *et al.* 2001 ; Gómez-Puertas *et al.* 1998). Il semble donc avéré que les anticorps produits contre le virus PPA permette de neutraliser une souche homologue mais ne permettent pas de lutter contre une infection avec un virus hétérologue.

Des études ont également été menées pour comprendre la réponse immunitaire cytotoxique, car il existe des preuves expérimentales de l'implication des lymphocytes T cytotoxiques dans la protection d'animaux infectés par des souches virulentes, mais non létales (Martins *et al.* 1993). On note cependant des différences en fonction de la souche virale utilisée pour l'infection. Même si on a démontré que la lyse des macrophages infectés était spécifiquement effectuée par les lymphocytes T CD8+ (Alonso *et al.* 1997), et que ces derniers produisaient de l'interféron gamma en quantité après leur mise en contact avec le virus, on ne connait que peu de protéines induisant une réponse immune cytotoxique. On a cependant mis en évidence que les protéines p32 et p72 étaient des cibles pour les lymphocytes T CD8+ (Alonso *et al.* 1997 ; Leitao *et al.* 2001 ; Leitao *et al.* 1998).

La réponse immune par la voie des NK, c'est-à-dire la réponse non spécifique, a également été explorée. Lors d'une infection virale, l'activité NK est largement augmentée. Les cellules NK s'attaquent aux cellules anormales dans l'organisme et les induits en apoptose puis lyse car elles présentent des antigènes du non-soi à leur surface. Une

première étude a été réalisée avec une souche portugaise peu virulente et non hémadsorbante de virus PPA, la souche ASF/NH/P68, dans le but de protéger les animaux contre une épreuve infectieuse avec la souche homologue virulente L60 (Martins & Leitao 1994). Dans cette étude, l'augmentation de l'activité NK induite par la « vaccination » avec la souche ASF/NH/P68 a été directement corrélée au niveau de protection obtenu contre la souche L60. Une seconde étude a confirmé que la réponse NK était directement à l'origine de la protection contre le virus PPA (Leitao *et al.* 2001). L'épreuve infectieuse avec la souche peu virulente ASF/NH/P68 a parfois induit une forme chronique de la maladie. Dans ce cas, l'activité NK chez les animaux infectés était restée au niveau de celle des animaux non infectés, c'est-à-dire très inférieure à celle des animaux résistant à l'infection.

Toutefois, l'activité NK induite par la souche ASF/NH/P68 résulterait de sa faible virulence probablement associée à la perte de ses propriétés hémadsorbantes et de la délétion d'une grande partie de son génome contenant les MGF. Or, tant le nombre de gènes appartenant à ces familles multigéniques que le phénomène d'hémadsorption ont été corrélés avec la virulence des souches (Afonso *et al.* 2004 ; Borca *et al.* 1998 ; Tabares *et al.* 1987). *A contrario*, une souche virulente ne disposant pas de ses caractéristiques a plutôt un effet inhibiteur de l'activité NK de cellules de sang périphérique de porc (Norley & Wardley 1983). Notons que d'autres protéines codées par le génome du virus PPA ont aussi été corrélées avec la virulence des souches. De façon non exhaustive, on peut citer la protéine j4R, une enzyme virale couplée à une ubiquitine cellulaire et ayant un rôle dans la régulation des voies de transcription des gènes (Goatley *et al.* 2002) ; la protéine A238L, inhibitrice du facteur de transcription NFκB de l'hôte (Yanez *et al.* 1995) et donc de l'induction des cytokines pro-inflammatoires nécessaires à la réponse immune innée (Powell *et al.* 1996 ; Revilla *et al.* 1998) ; ou encore des protéines anti-apoptotiques proches de celles de l'hôte, l'une se liant à la caspase 3 et inhibant son activité (Nogal *et al.* 2001) et l'autre liée à la famille des protéines Bcl-2, bien connues pour leur activité inhibitrice de l'apoptose (Revilla *et al.* 1997).

Ces protéines, qui participent à l'échappement du virus au système immunitaire de l'hôte sont fondamentales dans la virulence du virus. Au fil du temps, des virus de moindre virulence ont été isolés qui montraient de grandes délétions dans leur génome, particulièrement dans les régions codant pour les MGF, sans que ces dernières soient seules responsables de la pathogénicité des souches (Zsak *et al.* 1998 ; Zsak *et al.* 2001). Les différences dans la virulence, outre les signes cliniques associés, se caractérisent aussi par des taux de réplication virale plus ou moins élevés chez le porc domestique. Avec les souches très virulentes, le titre viral déterminé chez l'animal est supérieur à 10^8 TCID50, quand il est de 10^4 à 10^6 avec les souches modérément virulentes, et de seulement 10^2 à 10^3 pour les souches de faible virulence (Mebus *et al.* 1981). De même, chez les porcs sauvages africains, le titre viral peut varier au sein d'un même animal, de 10^2 TCID50 dans le sang circulant à $10^{6,6}$ dans les ganglions lymphatiques (Wilkinson 1989). Or, le taux de réplication a un impact sur le taux d'évolution du virus PPA. En effet, mathématiquement, un taux de

mutation bas peut cependant amener à une accumulation de diversité importante si le virus se réplique à haut niveau.

Sur le nombre total de gènes putatifs codés par le génome du virus PPA, environ 60 ont été estimés ne pas être essentiels à la réplication *in vitro* du virus (Yanez *et al.* 1995) et pourraient entrer dans la manipulation du système immunitaire de l'hôte. Combiné à une protection efficace contre le virus médiée par les anticorps (Neilan *et al.* 2004), cette subversion du système immunitaire de l'hôte expliquerait l'échec de toutes les stratégies vaccinales qui ont été tentées depuis près de 50 ans.

7-5- Epidémiologie – Hôtes – Transmission

Le virus PPA peut infecter l'ensemble des Suidae. Si les suidés sauvages africains sont considérés comme les premiers hôtes vertébrés du virus, les implications des diverses espèces ou sous-espèces de cette famille dans le cycle de la maladie ne sont pas toutes élucidées, principalement en ce qui concerne la transmission du virus aux porcs domestiques. Parmi les trois espèces majeures de suidés sauvages africains, le phacochère (*Phaecochoerus africanus*) est considéré comme le réservoir le plus important notamment à cause de son interaction avec les tiques molles du genre *Ornithodoros* dans lesquelles le virus peut également se répliquer (Haresnape *et al.* 1988 ; Penrith & Vosloo 2009 ; Thomson 1985 ; Wilkinson *et al.* 1988). De plus, la variété de son habitat lui permet d'être présent sur la totalité du continent africain sub-saharien. Cependant, le virus PPA infecte aussi les potamochères (*Potamochoerus porcus*) ainsi que les hylochères (*Hylochoerus meinertzhageni*). Quel que soit l'hôte suidé sauvage africain, l'infection est asymptomatique. En revanche, il est hautement létal chez le sanglier (*Sus scrofa*) dont les porcs domestiques sont presque tous des descendants.

La Peste porcine africaine est maintenue en Afrique via un cycle d'infections entre les phacochères et les tiques molles, espèces endophiles, qui infestent leur terrier (Plowright 1977 ; Thomson *et al.* 1980). Les tiques peuvent se contaminer de plusieurs manières : contamination par nourrissage sur des animaux infectés, puis transmission horizontale, verticale, et trans-stadiale entre elles (Plowright *et al.* 1974 ; Plowright *et al.* 1970a ; Plowright *et al.* 1970b). Enfin, elles inoculent le virus aux animaux lors du repas de sang : c'est le cycle selvatique (Figure 10) entretenu sur une période infinie (Plowright *et al.* 1969). Le virus PPA pourrait d'ailleurs être originellement un virus de tique qui se serait ensuite adapté aux suidés africains (Plowright 1977).

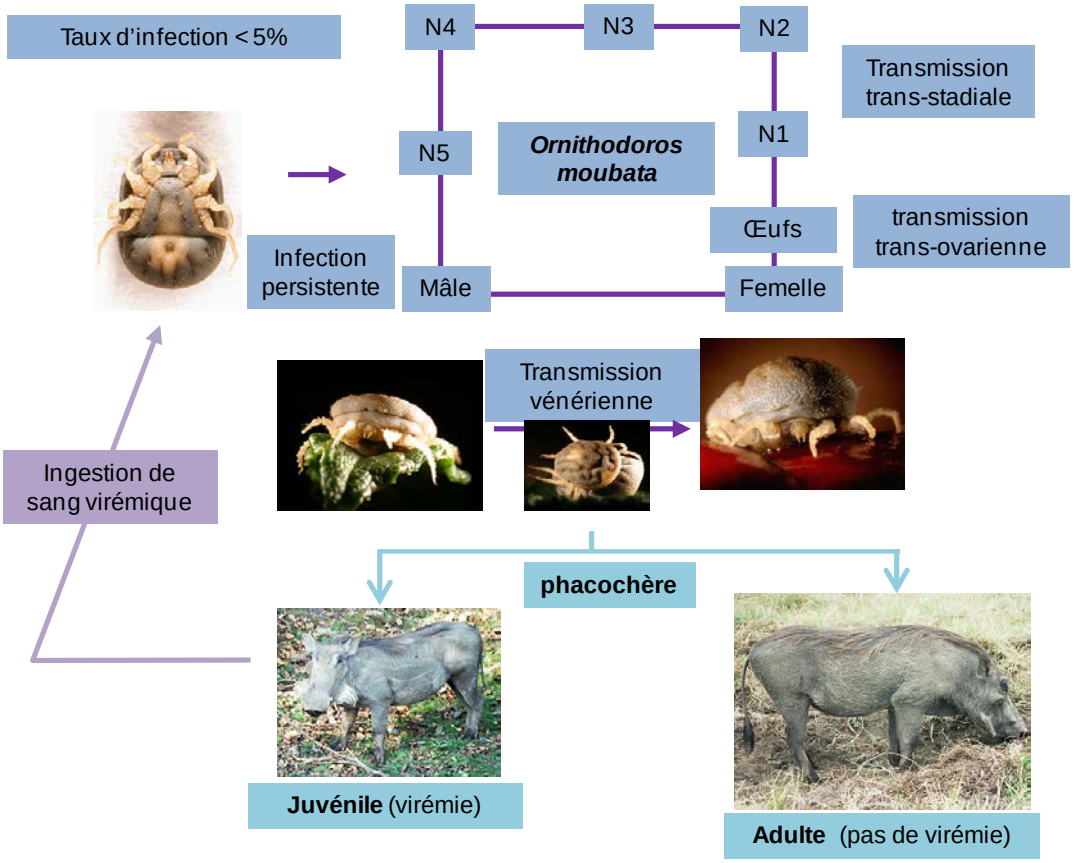


Figure 10 : Cycle selvatique impliquant les phacochères et les tiques. Le virus est présent à tous les stades de développement des tiques (transmission trans-stadiale) et la transmission entre tiques se fait aussi bien par voie sexuelle que trans-ovarienne. Les tiques vont ensuite infecter les suidés sauvages en se gorgeant sur eux ou s'infecter lors d'un repas de sang contaminé. On observe une virémie uniquement chez les jeunes suidés alors que le virus est présent dans les tissus des individus adultes.

Les voies de transmission du virus PPA aux porcs domestiques sont multiples. Tout d'abord par la voie du cycle sauvage. Jusqu'ici, aucune transmission directe par des phacochères infectés n'a pu être mise en évidence (Plowright 1981 ; Thomson *et al.* 1980). Le rôle épidémiologique des phacochères serait davantage dû aux tiques infectées que ces derniers peuvent transporter et amener jusqu'aux porcs en divagation, voire leur permettre d'établir des colonies dans les lieux d'élevage. Concernant les potamochères, en revanche, une étude expérimentale a montré que le virus pouvait être transmis aux porcs domestiques par contact, alors que l'inverse n'a pas été constaté (Anderson *et al.* 1998). Le rôle du potamochère dans le cycle selvatique du virus reste très mal élucidé car ne creusant pas de terrier, ses relations avec les tiques molles ne sont pas du même ordre que celles qui existent entre tiques et phacochères. Il a cependant été démontré expérimentalement que des potamochères infectés par le virus PPA pouvaient transmettre l'infection aux tiques lors du repas de sang. Son rôle dans l'épidémiologie de la maladie pourrait donc être non négligeable (Roger *et al.* 2001). De fait, considéré comme un nuisible pour les cultures auxquelles il s'attaque, les contacts avec les porcs en divagation sont plus que probables

(Haresnape 1984). De même, les contacts avec des carcasses de porcs morts de la maladie, qui restent contaminantes plusieurs jours, sont possibles car le potamochère peut se nourrir sur les carcasses (Penrith *et al.* 2004). Une dernière hypothèse concernant le rôle du potamochère consiste en la possibilité de contacts liés à la reproduction tels qu'avérés par l'existence d'hybrides avec le porc. Concernant l'hylochère, une seule preuve de portage de la maladie a été jusqu'ici mise en évidence, (Montgomery 1921), et son habitat strictement forestier rend son rôle épidémiologique très incertain.

Outre le cycle selvatique, il existe deux autres cycles de transmission de la PPA aux porcs domestiques. Le premier est dit enzootique intermédiaire et implique les tiques molles *O. porcinus porcinus* et *O. porcinus domesticus* associées aux élevages de porcs domestiques qui servent alors tant de réservoirs que de vecteurs du virus, et le second est dit domestique, car il ne fait intervenir que les porcs domestiques. Dans ce dernier cas, les voies de transmissions sont très nombreuses : contact, scarification, ingestion de déchets carnés contaminés, ou même simplement par contact avec des matières ou matériels contaminés. Ce cycle est à l'origine de foyers épidémiques qui peuvent s'étendre très rapidement, le virus étant très résistant et très persistant dans l'environnement.

7-6- Variabilité – Sérologie – Typage

La variabilité du virus PPA a été abordée sous deux aspects : antigénique et génomique. La comparaison des profils de restriction enzymatique des isolats a montré une grande variabilité, principalement au niveau des extrémités du génome (35 kb en 3' et 15 kb en 5') (Blasco *et al.* 1989a ; Blasco *et al.* 1989b ; Wesley *et al.* 1984). Cette grande variabilité n'a cependant pas permis d'établir des profils viraux types. L'on a simplement observé une variabilité supérieure des isolats en provenance de l'est et du sud de l'Afrique par rapport aux isolats européens, caribéens et sud américains. Sur le plan antigénique, il a été déterminé que 85% des protéines codées par le virus étaient identiques entre les différentes souches, les principales différences étant observées au sein des produits des MGF. Outre les indels présents dans les MGF, la variabilité en termes de protéines est portée par des différences dans le nombre d'acides aminés codés par des séquences répétées en tandem pour 14 protéinées virales, dont les protéines p54, α-like DNA-polymérase et CD2-homologue (Rodriguez *et al.* 1993a ; Yanez *et al.* 1995). Cette dernière est responsable de l'hémadsorption. Cette capacité à provoquer l'hémadsorption des globules rouges de porc a également été utilisée pour caractériser les isolats, car les anticorps générés par les porcs ayant survécu à une infection, ou ayant été infectés par des souches modérément virulentes inhibent l'hémadsorption de façon spécifique (Ruiz-Gonzalvo & Coll 1993). Cette première caractérisation avait permis de distinguer trois grands groupes de virus. Toutefois, des isolats non hémadsorbants ayant été par la suite isolés, cette méthode de classification a été abandonnée. La variabilité antigénique entre isolats a également été testée par le biais

d'anticorps monoclonaux dirigés contre des protéines structurales du virus (Whyard *et al.* 1985). Ceci a permis de classer 23 isolats testés en 6 groupes antigéniques, mais sans corrélation avec les régions d'origine des isolats. Cette méthode de classification a donc elle aussi été abandonnée.

8- Etat de l'art en phylogénie du virus PPA

La lutte la plus efficace contre la PPA, nous l'avons vu, passe par un diagnostic sûr et des mesures sanitaires drastiques. Or, l'efficacité des mesures zoosanitaires prises est soumise à la rapidité avec laquelle la souche responsable d'un foyer épidémique est identifiée. En effet, l'identification précise d'une souche virale permet de la tracer, c'est-à-dire de comprendre sa provenance, son entrée, et sa dispersion sur un territoire donné. Depuis l'échec des premières tentatives de classification des isolats de virus PPA, de nombreux outils de caractérisation moléculaire ont été développés, permettant une analyse plus fine de variations, même minimes, dans les séquences nucléotidiques. Néanmoins, les virus à ADN double brins ayant des taux de mutations et de substitutions très faibles, l'étude de ces variations devrait demander l'analyse de longs fragments pour être efficace. Chez les virus à ARN, l'analyse nucléotidique de gènes codant pour les protéines virales les plus immunogènes avait permis leur caractérisation et donc leur classification. Basées sur ce principe, des études ont montré que la capacité discriminante de certains gènes des virus à ADN, même très peu variables (Mbayed *et al.* 1998 ; Petrosillo *et al.* 2000), était comparable à celle du génome entier.

Ainsi, Bastos *et al*. (2003) ont réalisé la première étude en épidémiologie moléculaire portant sur la partie c-terminale du gène B6466L, le gène codant pour la protéine majeure de la capside virale (VP72). L'analyse a porté sur 58 isolats viraux, isolés à partir de porcs infectés au cours de foyers épidémiques ayant eu lieu en Europe, en Afrique (Ouest, Est et Sud), dans les Caraïbes ainsi qu'en Amérique du Sud. L'arbre phylogénétique généré a permis d'identifier 10 groupes distincts de virus, les génotypes I à X. Les isolats européens, ouest africains, caribéens et sud américains ont montré un lignage direct, avec une variabilité si faible qu'ils font tous partie du même génotype, le génotype I. Ainsi, malgré les 3 continents dans lesquels la maladie s'est introduite et a circulé pendant 50 ans, on n'observe que très peu de diversité entre isolats. Les neuf autres génotypes étaient formés d'isolats en provenance de l'est et du sud de l'Afrique et montraient d'avantage de diversité. Il était noté que ces isolats avaient été obtenus dans des régions où le cycle selvatique du virus était bien établi. En revanche aucun cycle sauvage n'a jamais été établi en Europe entre les tiques locales et les sangliers, même s'il a été montré que le virus PPA pouvait infecter les tiques molles du genre *O. erraticus* et les sangliers sauvages. Ceci corroborerait l'hypothèse que la diversité du virus serait générée par les vecteurs, réservoirs du virus (Bastos *et al.* 2003).

Cette étude a été confirmée et étendue deux ans plus tard, en 2005, par Lubisi (Lubisi *et al.* 2005). Cette étude comprenait cette fois 80 isolats viraux en provenance d'Afrique de l'Est et du Sud ainsi que 4 isolats en provenance d'Europe, d'Afrique de l'Ouest et d'Amérique du Sud. L'arbre induit des données de séquences déterminait à nouveau les 10 génotypes précédemment décrits, tout en observant six nouveaux génotypes (XI à XVI) en provenance d'Afrique de l'Est (Tanzanie, Malawi, Zambie). Le génotype I était détecté, cette fois, également dans le cycle selvatique, avec des virus isolés à partir de tiques et de potamochères sauvages. D'une façon générale, les virus isolés à partir de la faune sauvage montraient davantage de diversité que les souches isolées de porcs domestiques, laissant penser que le cycle domestique serait relativement indépendant du cycle selvatique et que l'introduction de nouveaux variants chez le porc domestique serait un évènement somme toute assez rare.

Le nombre de génotypes est passé à XXII après l'étude réalisée par Boshoff *et al.* (Boshoff *et al.* 2007). L'arbre inféré à partir de l'analyse de 42 isolats recouvrait les 16 génotypes déjà décrits, mais en déterminait six nouveaux par l'inclusion de nouvelles souches isolées lors de foyers épidémiques survenus dans le sud de l'Afrique durant la période 1973-1999. Cette étude montrait la présence de 3 lignages, dont deux ne comprenaient que des isolats en provenance de l'est ou du sud de l'Afrique. Outre la partie c-terminale du gène B646L, cette étude analysait également le gène B602L, situé dans la région centrale hypervariable du génome viral, et qui consiste en la répétition de tétramères d'acides aminés. Le gène B602L avait été utilisé pour la première fois en 2006 par Nix (Nix *et al.* 2006), en tentant de différencier les souches, par la détermination des différents tétramères qui le compose, et leur nombre. Ainsi, les 81 isolats étudiés, majoritairement isolés à partir de porcs domestiques et appartenant au génotype I avaient été discriminés en 31 groupes distincts. Le gène KP86R, codant pour une protéine transmembranaire (lui aussi composé de séquences répétées en tandem), ainsi que les régions intergéniques J286L et BtSj avaient également été analysés. Au travers de l'analyse de ces séquences, cette étude montrait 17 sous-groupes à l'intérieur du génotype I.

Certains génotypes le plus souvent très homogènes, tels que le VIII et le IX, semblent associés exclusivement aux cycles domestiques et intermédiaires, tandis que d'autres (V, X, XI, XII, XIII et XIV), sont également présents dans le cycle selvatique. Enfin, si certains génotypes semblent très géographiquement marqués (V, VI, IX, XI, XIII, XIV, XV et XVI), d'autres en revanche, circulent entre différentes régions (I, II, VIII, X et XII). Néanmoins, la co-circulation de souches nationales et transnationales au cours d'une même période de temps n'a pas permis de délimiter géographiquement la présence des souches.

Les gènes B646L et B602L ont été utilisés dans plusieurs études afin de tenter de discriminer les souches au niveau local (Gallardo *et al.* 2011a ; Gallardo *et al.* 2011b ; Lubisi *et al.* 2007 ; Owolodun *et al.* 2010) en se montrant d'un relatif intérêt. Depuis, d'autres gènes ont été utilisés pour tenter de discriminer les isolats au niveau régional, comme les

gènes E183L et CP204L (codant respectivement pour les protéines membranaires p54 et p32) (Gallardo *et al.* 2009 ; Giammarioli *et al.* 2011 ; Misinzo *et al.* 2010 ; Rowlands *et al.* 2008). Ces analyses, notamment celles utilisant le gène E183L ont tenté de discriminer des sous-génotypes à l'intérieur des vingt-deux génotypes connus, sans qu'il soit réellement possible de les déterminer.

Aucune investigation de l'évolution du virus n'a jusqu'ici été réalisée.

9- Nature et objectifs de la thèse

L'ajout de nouvelles souches virales dans les études phylogénétiques du virus PPA permet de déterminer un nombre croissant de génotypes viraux, passé de dix en 2003 à vingt-deux en 2007. L'échantillonnage tient donc une importance capitale dans la caractérisation des virus et donc dans la détermination des relations entre les souches. Or, la PPA est très inféodée à l'Afrique, région dans laquelle la collecte des échantillons, surtout ceux provenant de la faune sauvage, s'avère complexe, principalement en ce qui concerne la conservation et le transport des échantillons entre le lieu de prélèvement et le laboratoire. Jusqu'ici, les collectes étaient faites occasionnellement sur des organes ou du sang complet, dont la conservation et le transport demandaient une chaine de froid qu'il est souvent difficile de maintenir, ne serait-ce que pour des questions de coûts ou d'accès à un matériel adéquat. Ainsi, si les prélèvements en provenance de foyers épidémiques touchant des porcs domestiques d'élevage s'avèrent plus aisés, car directement effectués à l'abattoir, ceux en provenance de la faune sauvage ou de porcs en divagation restent limités en nombre, alors que leur apport en termes de compréhension de la diversité virale est sans doute fondamental. Dans ce travail, nous avons donc décidé dans un premier temps de développer un moyen efficace et peu coûteux de récolter, puis de transporter des prélèvements en s'affranchissant de la chaine du froid jusqu'au laboratoire pour permettre à la fois le diagnostic de l'infection et une caractérisation génétique du virus en cause. La méthode à développer devait être compatible avec la mise en œuvre de tests de diagnostic différents comme notamment la détection d'anticorps sériques par ELISA, l'isolement viral, ainsi que la PCR suivie d'un séquençage dans le but de détecter et d'étudier au moyen de la phylogénie, l'émergence de nouvelles souches ou la réémergence d'anciennes souches virales.

Dans un second temps, les outils d'étude phylogénétique tels que connus jusqu'alors ont été mis en œuvre pour évaluer leur pertinence en terme de traçabilité épidémiologique. A ce titre, nous avons pu caractériser un nombre important de souches circulant à Madagascar et faire le lien avec l'introduction du virus dans le Caucase en 2007.

Enfin, nous nous sommes intéressés à reconsidérer la classification du virus PPA, enracinée dans le groupe des grands virus à ADN nucléocytoplasmiques, en utilisant des méthodes modernes de reconstructions phylogénétiques, puis avons évalué la pertinence de

ces méthodes et de nos jeux de données pour dater l'origine du virus PPA et de ses différents clades.

Partie 1

Mise au point d'une méthode simple adaptée à la détection et à la caractérisation du virus PPA dans les conditions tropicales

Le berceau du virus PPA est en Afrique de l'est. C'est un virus qui est détecté depuis plus d'un siècle sur ce continent et qui est largement répandu. Pour autant, l'obtention d'isolats caractérisés sur le plan génétique est confrontée à la difficulté d'accéder aux animaux infectés, de prélever et de conserver de manière satisfaisante les échantillons biologiques jusqu'à leur acheminement au laboratoire et enfin de mettre en œuvre un diagnostic simple et peu couteux. Dans le but d'essayer de répondre à cet enjeu, nous nous sommes intéressés à imaginer une méthode qui permettrait de répondre à l'ensemble de ces difficultés. Après avoir analysé la littérature sur les questions de prélèvements et de conservation longue durée d'échantillons biologiques hors chaine du froid, nous avons décidé d'explorer l'utilisation du papier buvard. Pour simplifier et réduire le coût du diagnostic de laboratoire, notamment par PCR, nous avons évalué la possibilité d'utiliser les papiers buvards directement dans les tubes diagnostic, sans passer par une extraction préalable des acides nucléiques. Enfin, nous avons validé le fait qu'à partir des papiers buvards, nous restions en capacité de caractériser génétiquement les virus en cause.

Cette stratégie a été conduite à son terme et a permis la publication d'un premier article scientifique reproduit dans la partie qui suit.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Virological Methods 146 (2007) 257–265
Journal of Virological Methods
www.elsevier.com/locate/jviromet

Long-term storage at tropical temperature of dried-blood filter papers for detection and genotyping of RNA and DNA viruses by direct PCR

V. Michaud [a], P. Gil [a], O. Kwiatek [a], S. Prome [a], L. Dixon [b], L. Romero [c], M.-F. Le Potier [d], M. Arias [c], E. Couacy-Hymann [e], F. Roger [f], G. Libeau [a], E. Albina [a,*]

[a] *CIRAD, UR Contrôle des Maladies, Montpellier F-34398, France*
[b] *Institute for Animal Health, Ash Road, Pirbright, Woking, Surrey GU24 ONF, United Kingdom*
[c] *Centro de Investigacion en Sanidad Animal, Ctra Valdeolmo-El Casar, 28130 Madrid, Spain*
[d] *AFSSA, Laboratoire d'études et de recherches avicoles et porcines, Ploufragan F-22 440, France*
[e] *Laboratoire Central de Pathologie Animale, BP 206 Bingerville, Côte-d'Ivoire*
[f] *CIRAD, UR Epidémiologie, Montpellier F-34398, France*

Received 7 December 2006; received in revised form 6 July 2007; accepted 10 July 2007
Available online 21 August 2007

Abstract

In tropical countries the diagnosis of viral infections of humans or animals is often hampered by the lack of suitable clinical material and the necessity to maintain a cold chain for sample preservation up to the laboratory. This study describes the use of filter papers for rapid sample collection, and the molecular detection and genotyping of viruses when stored over long periods at elevated temperatures. Infected blood was collected on filter papers, dried and stored at different temperatures (22, 32 and 37 °C) for various periods (up to 9 months). Two animal viruses, African swine fever, a large double-stranded DNA virus and Peste des Petits Ruminants, a negative single-stranded RNA virus, were used to validate the method. Filter papers with dried blood containing virus or control plasmid DNA were cut in small 5 mm^2 pieces and added directly to the PCR tube for conventional PCR. Nucleic acid from both viruses could still be detected after 3 months at 32 °C. Moreover, the DNA virus could be detected at least 9 months after conservation at 37 °C. PCR products obtained from the filter papers were sequenced and phylogenetic analysis carried out. The results were consistent with published sequences, demonstrating that this method can be used for virus genotyping.

Keywords: Filter paper; PCR; Genetic analysis; Phylogeny; African swine fever; Peste des petits ruminants

1. Introduction

Rapid diagnosis of infections is essential for the control of diseases. Not only the detection of the infection but also genotyping of the causative agent is often needed for efficient disease surveillance and control programmes (Zollner, 2004). A reliable procedure is required for sample collection, preservation, transport to laboratories and testing. Generally, samples are preserved using ice packs, dry ice or liquid nitrogen according to the nature of the agent and the expected delay in shipment. However, in some regions, it may be difficult to maintain a cold chain and transport to the laboratory may take more than 3–4 days. The situation is complicated by the international regulations regarding the transport of frozen forms of biohazardous agents (liquid or tissue) surrounded by dry ice or liquid nitrogen. In recent years, several studies have demonstrated the potential interest of filter papers for the collection and storage of biological materials. Using this, the cold chain is not necessary and thus the transport under international regulations is rendered easier. Filter papers have been widely used for blood preservation and antibody detection in the laboratory for human and animal diseases (Behets et al., 1992; De Swart et al., 2001; Hogrefe et al., 2002; Hutet et al., 2003; De la C Herrera et al., 2006). In addition, they have been repeatedly used to detect the genomes of DNA or RNA viruses by PCR (Spagnuolo-Weaver et al., 1998; Hattermann et al., 2002). Filter papers can also be used for the study of genetic variability of viruses (Pitcovski et al., 1999). Different types of filter papers

* Corresponding author at: CIRAD, UR Contrôle des Maladies, TA 30/G, Campus International de Baillarguet, 34398 Montpellier cedex 5, France. Tel.: +33 4 67 59 37 05; fax: +33 4 67 59 37 98.
E-mail address: emmanuel.albina@cirad.fr (E. Albina).

0166-0934/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jviromet.2007.07.006

have been tested, some have no additives, whereas others are specifically designed for the preservation of genomic material and are impregnated with chemicals that lyse cells and denature proteins (Natarajan et al., 2000). The virus genome can be detected after extraction of the genomic material (Prado et al., 2005; Zhou et al., 2006) or by direct PCR without extraction (Yourno and Conroy, 1992; Pitcovski et al., 1999; Kailash et al., 2002). Filter papers have been shown to be suitable for the conservation of either DNA or RNA viruses for extended periods of time (up to 4–11 years) at moderate or tropical temperatures (Li et al., 2004; Chaisomchit et al., 2005). However, there is no study showing that conventional filter papers can be used for long-term conservation of DNA and RNA viruses at tropical temperature and detection by direct PCR followed by genetic characterisation of the viruses. Two different animal viruses were used as models for the validation of the procedure. Both are responsible of two highly contagious and fatal diseases that are listed by the World Animal Health Organisation (OIE) among the 15 most serious diseases. African swine fever virus (ASFV) is a double-strand DNA enveloped virus that belongs to the family *Asfarviridae*, genus *Asfivirus*. The causative agent of peste des petits ruminants (PPR) is a single-strand negative RNA enveloped virus which is a member of the genus *Morbillivirus* within the family *Paramyxoviridae*. Both viruses were used in this study to determine the threshold for detecting the virus genomes by direct PCR from filter papers. The sensitivity of the method was compared to a conventional diagnostic test and the effect of long-term storage at high temperature was assessed.

2. Material and methods

2.1. Samples and reagents

Pigs free of the specific pathogens listed by the OIE (SPF pigs) were bled and filter papers impregnated with their blood and dried. Filter papers containing dried blood were prepared in the same way from African farmed goats. These goats were confirmed by serology as free of PPR and rinderpest (RP) infections twice. These filter papers served as negative controls in the PCR tests and also as supports for serial dilutions of each virus or plasmid containing the target gene in order to determine the detection limit of the technique. Additionally, 80 pigs and 73 goats from French farms were sampled on filter papers and tested in PCR to assess the specificity of the method. Four pigs, confirmed to be free of ASF infection by antibody detection, were infected experimentally by the intramuscular route with a Spanish strain of ASFV isolated in 1970 (strain E70, Sierra et al., 1990). Two other pigs were challenged with the same strain but by the oronasal route and placed in direct contact with one susceptible pig. Four additional pigs were inoculated by the oronasal route with the Lisbon 60 (L60) ASFV strain isolated in 1960 (Leitao et al., 2001). In this last experiment, four non-infected pigs were kept as negative controls. Blood was collected both in EDTA tubes and on filter papers, on the day of challenge and thereafter at different days post-challenge as indicated in results.

African dwarf goats aged between 2 and 3 years which were shown to be free of PPR and RP antibodies were experimentally infected with four different strains of PPRV corresponding to different lineages (Couacy-Hymann et al., 2005): strains Nigeria 75–1 (Nig/75–1, lineage 1), Côte d'Ivoire 89 (CI89, lineage 2), Ethiopia 94 (Ethio94, lineage 3) and India-Calcutta (Calcutta, lineage 4). Each strain was inoculated subcutaneously in one goat. Blood samples and filter papers were collected at different time post-challenge as indicated in results.

As positive control of PCR and also for detection limit purposes, the target genes of both viruses were cloned in plasmids. The VP72/73 and N gene of ASFV and PPRV, respectively, were selected as target genes as recommended by the Manual of Standards and Diagnostics of the World Animal Health Organisation (OIE). Plasmids were produced in large quantities using the Maxiprep Kit (Qiagen, France) according to the manufacturer's instructions. The DNA concentration was determined by UV absorbance. The number of gene copies in the plasmid preparation was estimated as a relation of the concentration established by UV absorbance and the plasmid size using the following formula 1.0 A_{260} unit dsDNA = 50 μg/ml and 1 μg of 1000 bp DNA = 9.1×10^{11} molecules. For ASFV, 10-fold serial dilutions of the plasmid were prepared with the objective to have copy numbers corresponding to the virus titres. For PPRV, the target N gene of PPRV was cloned under control of the bacteriophage T7 polymerase promoter and transcribed *in vitro* using the T7 RiboMAX Express Large Scale RNA Production System kit (Promega, France) according to the manufacturer's instructions. The number of RNA copies were then estimated as a factor of concentration, as determined by UV absorbance and RNA molecule size by using the formula 1.0 A_{260} unit ssRNA = 40 μg/ml and 1 μg of 1000 b RNA = 18×10^{11} molecules. Both plasmid (ASFV) and RNA (PPRV) serial dilutions were spotted on the filter papers to estimate the detection limit of the method.

A strain of ASFV (BA71V strain isolated in Spain in 1971 and adapted to grow on Vero cells) and the vaccine PPRV strain Nigeria 75/1 (attenuated by serial passages on Vero cells) were amplified and titrated according to the method of Kaerber (1931). Titres were expressed as tissue culture infectious doses 50% ($TCID_{50}$) per ml. Serial dilutions of these viruses were spotted on filter papers and used for the determination of detection limit of the method.

2.2. Filter paper preparation

The Whatman 3MM filter paper (VWR, Fontenay-sous-Bois, France), often used for storage and detection of genetic or protein materials, was primarily selected for this study because of its low cost. However, Whatman FTA cards (Dutcher, Brumath, France), specifically designed for nucleic acid stabilization, were used to compare the detection limit of ASFV.

In order to prepare calibration standards, 5 mm^2 surfaces of filter papers containing dried blood collected from SPF pigs or susceptible goats were impregnated with 2 μl of 1/10 serial dilutions of either plasmids containing the ASFV or PPRV target genes, *in vitro* transcribed RNA from PPRV target gene or

titrated viruses. To compare direct spotting of diluted virus on blood-dried filter papers or spotting of virus diluted first in blood and spotted on filter papers, ASFV and PPRV were serially diluted either in cell culture medium (EMEM, Eurobio, France) or in the blood of one non-infected pig and goat, respectively. Dilutions of the virus in medium were spotted on blood-dried filter papers as previously described. Dilutions in blood were incubated at 37 °C for 30 min and then spotted on filter papers, allowed to dry and tested in the direct PCR.

Bloods from farm pigs and goats were deposited on filter papers, allow to dry and stored at −80 °C until use. Experimentally infected pigs or goats were bled and filter paper strips were immediately impregnated and allow to dry. Once dried, the strips were stored at −80 °C until use. Filter papers found positive by ASFV PCR were stored for 9 months at 22–25 °C or 37 °C. Those found positive by PPRV PCR were stored for 3 months at 32 °C. All filter papers were stored in an environment with 50–70% of humidity. These filter papers were tested once a month.

2.3. PCR

Filter papers containing dried blood from infected pigs were directly processed into the PCR tubes without any nucleic acid extraction. Pieces of 5 mm^2 were placed into 0.2 ml PCR tubes. Reaction mix was added to a final volume of 80 μl to allow proper soaking of filter papers. For DNA extraction from blood collected on EDTA tubes, 100 μl of whole blood were treated with the DNeasy kit (Qiagen) according to the manufacturer's instructions. PCR was then run with 5 μl of extracted DNA in a final volume of 50 μl.

Different primer pairs were initially designed and tested on filter papers and the best pair was selected for further testing (data not shown). The proof reading polymerase (Taq pol Pfu, Stratagene, Amsterdam) was used to allow direct sequencing of the PCR products. For ASFV detection, the reaction mix consisted of 0.4 μM of each primer [forward: 5′-TCggAgATgTTCCAggTAgg-3′, reverse: 5′-CgCAAAAggATTTggTgAAT-3′], 250 μM dNTP, 2.5 units of Pfu polymerase. After amplification (5 min at 95 °C, then 35 cycles, 30 s at 95 °C, 30 s at 55 °C and 30 s at 72 °C, and finally 7 min at 72 °C), a DNA fragment of 346 base pairs was visualized on agarose gels. However, to ensure that the most sensitive PCR methodology was used, two other polymerases and PCR protocols for ASFV detection were also evaluated. One uses the master mix of Eppendorf (Dutcher) which has the advantage that is ready to use. The other is the hotstart immolase DNA polymerase (Bioline, Abcys, France) and was used by Basto et al. (2006) to detect ASFV in ticks. In all PCR runs, a negative control consisting of a dried-blood filter paper from a SPF pig was included. DNA from ASFV-infected cell culture was the positive control. This DNA was extracted with the DNeasy kit (Qiagen).

A one step RT-PCR (Qiagen) was performed on a single punched disc from infected goats, without any extraction of the viral RNA. The primer set NP3/NP4 (Couacy-Hymann et al., 2002), used in the study was designed from the nucleoprotein gene sequences to amplify specifically PPRV. Pieces of filter paper of 5 mm^2 were placed into 0.2 ml PCR tubes and 33 μl of RNase free water were added. The tubes were heated at 95 °C for 10 min and then immediately placed on ice. Seventeen microlitre of reaction mix consisting of 10× PCR reaction buffer, 0.6 μM of each primer [forward: 5′-gTCTCggAAATCgCCTCACAgACT-3, reverse: 5′-CCTCCTCCTggTCCT CCAgAATCT-3′], 400 μM of each dNTP and 2 μl of the Qiagen OneStep RT-PCR Enzyme Mix (Qiagen). After reverse transcription and amplification (30 min at 50 °C, 15 min at 95 °C then 35 cycles, 30 s at 95 °C, 30 s at 53 °C and 30 s at 72 °C, and finally 10 min at 72 °C), DNA bands of the expected 352 bp size were obtained. In the PCR reactions, a negative control was included, which consisted of RNA extracted from non-infected cell cultures with the "Nucleospin RNA virus" kit (Macherey Nagel, France).

2.4. Sequencing and phylogenetic analysis

Phylogenetic analysis was carried out on the sequences obtained from PCR products amplified from filter papers collected from a pig infected with the Lisbon 60 (L60) ASFV strain and from 2 goats, one infected with the PPRV Côte d'Ivoire 1989 (CI89) and the other with the Ethiopian strain (Ethio94). PCR products amplified from filter papers were purified from the gels using the "Qiaquick PCR purification" kit (Qiagen) and directly sequenced by GATC (Germany) using the same PCR primers. Analysis of sequences was performed using Vector NTI-9 package (Invitrogen, USA). Multiple alignments of sequences were done with the Clustal Wallis application included in the Vector NTI package. Sequences were retrieved from Genbank (Table 1) or generated in this study (Michaud et al., unpublished). The alignments were exported in msf file format for conversion in a Phylip 3.2 format using Bioedit software to allow phylogenetic analysis (Hall, 1999). Phylogenetic analysis was carried out using the neighbor-joining method based on the principle of parsimony (Saitou and Nei, 1987), included in the Darwin5 software (Perrier et al., 2003). Dissimilarities and distances between the sequences were first determined by Darwin5 and trees were generated with the TreeCon MATRIXW program (Van de Peer and De Wachter, 1993) included in Darwin5. Tree construction was based on the unweighted neighbor-joining method proposed by Gascuel (1997). Bootstraps were determined on 1000 replicates.

3. Results

3.1. Determination of the detection limit of PCR on filter papers containing dried blood

The limits for detection of virus nucleic acid by PCR amplification from filter papers containing dried blood were determined using serial dilutions of either ASFV or PPRV viruses or corresponding plasmids containing the target gene. The use of FTA cards instead of Whatman 3MM with Taq pfu did not improve the analytical sensitivity of the method (data not shown). The Taq polymerase Bioline, when use for the detection of ASFV

Table 1
List of sequences retrieved from Genbank and used for the phylogenetic analysis

Name of virus isolate	Country of origin	Date of isolation	Reference or year of submission to GenBank	GenBank accession nos
African swine fever virus				
DRC/67	DRC	1967	Zsak et al. submitted 2004	AY578708
Ba71v	Spain	1971	Lopez-Otin et al., 1990	M34142
DR-2	Dominican Republic	?	Yu et al., 1996	L76727
L60	Portugal	1960	Bastos et al., 2003	AF301539
CAM/82	Cameroon	1982	Bastos et al., 2003	AF301544
KEN/64	Kenya	1964	Zsak et al. submitted 2004	AY578697
Haiti/79	Haiti	1979	Zsak et al. submitted 2004	AY578695
E75	Spain	1975	Zsak et al. submitted 2004	AY578693
Mkuzi/79	RSA	1979	Kutish and Rock submitted 2003	AY261362
ZIM/83	Zimbabwe	1983	Zsak et al. submitted 2004	AY578705
Toliara/98	Madagascar	1998	Michaud et al, unpublished	DQ875934
Moronda/02	Madagascar	2002	Michaud et al, unpublished	DQ875935
RSA/96-5	RSA (Noord Biabant)	1996	Zsak et al. submitted 2004	AY578701
RSA/96-4	RSA (Wildebeeslagte)	1996	Zsak et al. submitted 2004	AY578699
RSA/96-3	RSA (Fairfield)	1996	Zsak et al. submitted 2004	AY578696
RSA/96-2	RSA (Nooitverwacht)	1996	Zsak et al. submitted 2004	AY578694
NAM-Warth	Namibia	?	Kutish and Rock submitted 2003	AY261366.
Tengani/62	Malawi	1960	Bastos et al., 2003	AF301541
RSA-Warm	RSA	?	Kutish and Rock submitted 2003	AY261365
RSA/96-1	RSA (Crocodile)	1996	Zsak et al. submitted 2004	AY578691
RSA/96-6	RSA (Pretorisuskop)	1996	Kutish and Rock submitted 2003	AY261363
KEN/50	Kenya	1950	Kutish and Rock submitted 2003	AY261360
UGA/65	Uganda	1965	Yu et al., 1996	L27499
Lil20/1	Malawi	?	Yozawa et al., 1994	U03762
Peste des petits ruminants virus				
Nig75/1	Nigeria	1975	Kwiatek et al., 2007	DQ840160
Nig76/1	Nigeria	1976	Kwiatek et al., 2007	DQ840164
Ghana78	Ghana	1978	Kwiatek et al., 2007	DQ840166
Nig75/3	Nigeria	1975	Kwiatek et al., 2007	DQ840162
Nig75/2	Nigeria	1975	Kwiatek et al., 2007	DQ840161
Mali1	Mali	1999	Kwiatek et al., 2007	DQ840192
Iran98	Iran	1998	Kwiatek et al., 2007	DQ840185
Saoudi/7	Saudi Arabia	1999	Kwiatek et al., 2007	DQ840195
Saoudi/8	Saudi Arabia	1999	Kwiatek et al., 2007	DQ840197
Turkey96	Turkey	1996	Kwiatek et al., 2007	DQ840184
Israel95/3	Israel	1995	Kwiatek et al., 2007	DQ840181
Israel/2	Israel	1998	Kwiatek et al., 2007	DQ840178
Israel	Israel	1993	Kwiatek et al., 2007	DQ840173
Iran/3	Iran	1994	Kwiatek et al., 2007	DQ840186
India94	India	1994	Kwiatek et al., 2007	DQ840176
Calcutta	India	1995	Kwiatek et al., 2007	DQ840177
Oman83/2	Oman	1983	Kwiatek et al., 2007	DQ840168
UAE86	United Arab Emirats	1986	Kwiatek et al., 2007	DQ840169
Sudan72	Sudan	1972	Kwiatek et al., 2007	DQ840158
Ethio96	Ethiopia	1996	Kwiatek et al., 2007	DQ840183
Ethio94	Ethiopia	1994	Kwiatek et al., 2007	DQ840175
Guinea88	Guinea	1988	Kwiatek et al., 2007	DQ840170
CI89	Ivory coast	1989	Kwiatek et al., 2007	DQ840199
Burki88	Burkina Faso	1988	Kwiatek et al., 2007	DQ840172
Bissau89	Guinea-Bissau	1989	Kwiatek et al., 2007	DQ840171
Seneg94	Senegal	1994	Kwiatek et al., 2007	DQ840174
Seneg68	Senegal	1968	Kwiatek et al., 2007	DQ840165

in ticks was completely inefficient on dried-blood filter papers either collected on Whatman 3MM or FTA cards. In contrast, the Eppendorf master mix compared to Taq pfu, gave a lower sensitivity on Whatman 3MM but a higher sensitivity on FTA cards. Since FTA cards were more expensive and required extra-steps of washings before running the PCR and since a secondary aim of this project was also to sequence the amplified product, it was therefore decided to use Whatman 3MM and Taq pfu to avoid errors during amplification.

Results for the selected protocol using Taq pfu to amplify target genes spotted as either plasmid, RNA transcript or virus on Whatman 3MM filter papers are shown in Fig. 1. For ASFV, PCR from filter papers detected less than two copies of the VP72 gene and less than one $TCID_{50}$. The detection limit for PPRV was

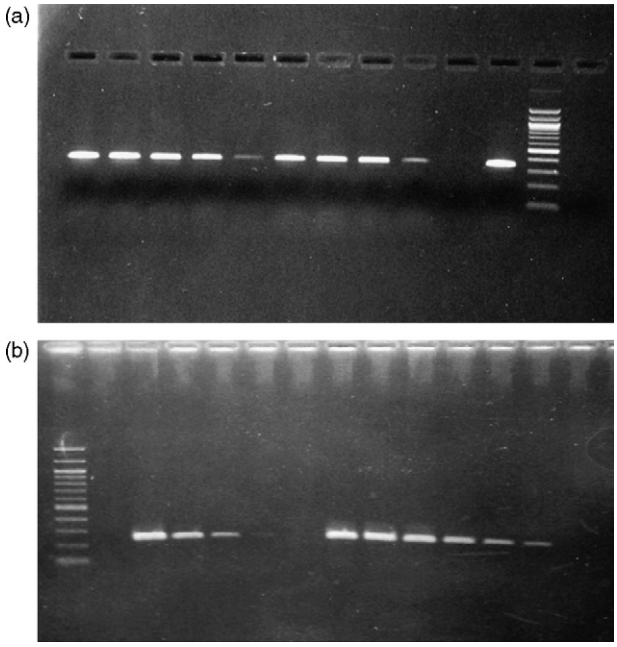


Fig. 1. Determination of the detection limit of direct PCR on blood-dried filter papers. This figure show one of two repetitions of the assay (see results section). (a) Detection of ASFV DNA virus spotted on FP as serial dilutions of either $TCID_{50}$ or plasmid DNA copies containing the VP72 virus gene. Lanes 1 to 5 are 10-fold serial dilutions of ASFV, from 1260 to 0.126 $TCID_{50}$. Lanes 6 to 10 are 10-fold serial dilutions of VP72 plasmid, from 1260 to 0.1260 gene copies. Lanes 11, 12 and 13 are control + (ASFV extracted from infected cell cultures), ladder and control—(DNA extracted from non-infected cell cultures), respectively. (b) Detection of PPRV RNA virus as dilutions of either $TCID_{50}$ or RNA transcript copies from the N gene. Lane 1 is the ladder and lanes 2 and 7 are negative controls (RNA extracted from non-infected cell cultures). Lanes 3–6 are 10-fold serial dilutions of PPRV, from 398 to 0.398 $TCID_{50}$. Lanes 8 to 14 are 10-fold serial dilutions of PPRV N transcripts, from 2×10^9 to 2×10^3.

2×10^4 copies of N-RNA transcript and less than one $TCID_{50}$ (Fig. 1). Nonetheless, the PCR from filter papers was able to detect at least one copy of the N gene cloned into a plasmid (data not shown). The determination of the detection limit for the two viruses was repeated and the method could detect at least two $TCID_{50}$ of each virus spotted on blood-dried filter papers. When the viruses were first diluted in the blood and incubated for 30 min at +37 °C before spotting on the filter papers, there was no loss of sensitivity as compared with the first procedure.

3.2. Application of PCR on filter papers collected from farms or experimentally infected pigs and goats

The direct PCR carried out on blood-dried filter papers collected on 80 farm pigs and 73 farm goats gave a negative result as expected which gives an estimated specificity higher than 96% on the two separate populations and higher than 98% on the whole population. ASFV DNA in blood of infected pigs was detected by PCR in parallel from blood collected in EDTA tubes and dried on filter papers. The results are shown in Table 2. All samples were found negative on the day of the challenge. Three days post-challenge in trial 1, all filter papers were positive by PCR and a good agreement was observed with the results from the samples from EDTA tubes. In trial 2, the results were very similar except that fewer samples were found positive on day 3. All non-infected pigs in trial 2 were found negative both for filter papers and EDTA tubes all over the experimental period. The percentage agreement between PCR from samples collected in EDTA tubes and filter papers (Kappa coefficient) was 89.5%, which is satisfactory (Jakobsson and Westergren, 2005). For both ASFV and PPRV infected animals, the kinetics of virus genome detection in peripheral blood was established (Table 3). Virus was detected from the two infected pigs with the same kinetics, starting from day 8 and lasting for at least 17 days post-challenge. The third pig, placed in direct contact with the two others, developed a DNAemia almost at the same time, suggesting that this animal had been infected at the same time as the other, presumably by inoculum discharges. However, the virus dose received by this contact pig was probably reduced since the DNAemia did not last after day 12 post-challenge. RNAemia in the infected goats was variable according to the virulence of the strain. Infected goats 2 and 3 were infected by highly virulent strains and died at day nine post-challenge. They had PPRV RNA in their blood as soon as day 4 or 5 after the oronasal challenge. In contrast, the two other goats survived the infection and PPRV RNA was detected in their blood, 6 or 9 days after challenge and thereafter.

Table 2
Results of ASFV DNA detection in blood of pigs infected by the intramuscular route with the strain E70 and by the oronasal route with strain L60

Days post-challenge	0		3		5–6		7	
	Filter papers	EDTA blood	Filter papers	EDTA blood	Filter papers	EDTA blood	Filter papers	EDTA blood
Trial 1 (strain E70 intramuscular)								
Pig 1	–	–	+	+	ns	–	ns	ns
Pig 2	–	–	+	ns	+	ns	ns	ns
Pig 3	–	–	+	+	+	ns	ns	ns
Pig 4	–	–	+	+	–	ns	ns	ns
Trial 2 (strain L60 oronasal)								
Pig 5	–	–	–	–	+	+	+	–
Pig 6	–	–	–	+	+	+	dead	dead
Pig7	–	–	–	ns	+	+	dead	dead
Pig 8	–	–	+	+	+	+	dead	dead

Pigs were bled in parallel on EDTA tubes and on filter papers.
ns: not sampled.

Table 3
Results of the detection of ASFV and PPRV genome in blood collected on filter papers from pigs or goats infected experimentally by the oronasal route

Days post-infection	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Pig challenge trial																
Infected pig 1	–	–	ns	ns	–	–	+	+	+	+	ns	ns	+	+	+	+
Infected pig 2	–	–	ns	ns	–	–	+	+	+	+	ns	ns	+	+	+	+
Contact pig	–	–	ns	ns	–	–	–	+	+	+	ns	ns	–	–	–	–
Goat challenge trial																
Infected goat 1	–	–	–	–	–	–	–	+	+	+	+	–	–	–	ns	ns
Infected goat 2	–	–	+	+	+	+	–	dead								
Infected goat 3	–	–	–	+	+	+	+	dead								
Infected goat 4	–	–	–	–	+	+	–	–	–	–	–	–	–	–	ns	ns

ns = not sampled.
ASFV strain was the E70. Goats 1 to 4 were infected by PPRV strains Nigeria 75–1 (Nig/75-1, lineage 1, low virulent), Ethiopia 94 (Ethio94, lineage 3, high virulent), Côte d'Ivoire 89 (CI89, lineage 2, high virulent) and India-Calcutta (Calcutta, lineage 4, low virulent), respectively.

3.3. Long-term storage of blood-dried filter papers at high temperatures

Filter papers initially scored positive for ASFV by PCR were still detected as positive after storage for 9 months at 22–25 °C or 37 °C. Filter papers positive for PPRV by PCR were also still positive after 3 months at 32 °C. No loss of sensitivity was observed after this period. Negative filter papers remained negative.

3.4. Genotyping of the strains collected on filter papers

After PCR, the amplicons were purified and sent for sequencing. The ASFV sequence obtained from filter papers was identical to the Lisbon 60 sequence deposited on Genbank (accession no AF449480). The sequence of PPRV strain Ethio94 was identical to the one previously established (Roeder et al., 1994). The other sequence obtained for strain CI89 differed by only one nucleotide. It is believed that the initial sequence determined previously was incorrect at this position (Diallo, personal communication) since a pyrimidine-to-purine transversion (U → G) was observed whereas the other 27 strains sequenced in this study have a G at this position. The sequences generated from filter papers were aligned with other established sequences and phylogenetic trees were generated as described in material and methods. Fig. 2 shows the result for one strain of ASFV and two strains of PPRV belonging to two different lineages. As expected, the ASFV sequence clustered within the group I consisting of European, West African and South American strains. The maximum number of variable positions observed in this 346 bp region was 29 (11.9%). The resulting phylogenetic tree was almost identical to the one produced by Bastos et al. (2003) and the positions of the groups in Fig. 2a was directly derived from their work. Interestingly, the partial sequence of the VP72 gene used in this phylogenetic analysis was not in the same region than the one used by Bastos et al. (2003), thus illustrating that different regions of the VP72 gene may be used for the phylogenetic analysis of ASFV strains. Similarly, CI89 and Ethio94 strains sequenced from PCR products containing the nucleoprotein gene amplified from the filter papers were clustered into lineages 2 and 3 as already established by Dhar et al. (2002) on partial sequence of the fusion protein gene and latter on confirmed by our group on partial sequence of the N gene (Kwiatek et al., 2007). The maximum nucleotide divergence observed in the 255 bp region was 52 (20%).

4. Discussion

Filter papers containing dried blood are interesting sampling systems for the conservation of biological materials when the use of cold chain is impracticable, for instance under extreme climatic conditions encountered in Africa. They can serve for many diagnostic purposes, including the detection of serum proteins like antibodies, genetic material like somatic DNA for the diagnosis of genetic diseases and virus or parasite genomes. A relatively high number of viruses have already been detected on filter papers by PCR among which, the Human immunodeficiency virus (Yourno and Conroy, 1992; Beck et al., 2001), Measles virus (De Swart et al., 2001; Katz et al., 2002; Mosquera et al., 2004), Hepatitis C virus (Abe and Konomi, 1998), dengue virus (Prado et al., 2005), Human papillomavirus (Kailash et al., 2002) and various animal DNA viruses (Hattermann et al., 2002; Wang et al., 2002; Guy-Gonzague et al., 2003) and RNA viruses (Spagnuolo-Weaver et al., 1998; Moscoso et al., 2005; Dubay et al., 2006). This illustrates that sample collection and storage on filter papers can be adapted to a wide range of RNA and DNA viruses. Generally, the use of filter papers requires the pre-treatment of the material in order to extract the biomolecules to be detected and/or to be sequenced. In this study, filter papers were used after long-term storage at high temperatures in a direct PCR test without any previous extraction of nucleic acids. Whatman 3MM filter papers used in this study are cheap and although they are not specifically designed for nucleic acids preservation, they proved to be efficient in this study and others (Kailash et al., 2002). In other publications, FTA Whatman cards were used, which allow cell lysis and the binding of nucleic acids (Beck et al., 2001; Moscoso et al., 2005). These FTA cards have been shown to preserve genetic materials for extended periods of time: 4 years at 22–24 °C (Li et al., 2004) and up to 11 years at ambient tropical conditions (Chaisomchit et al., 2005). However, FTA cards and elution buffer are expensive compared to Whatman 3MM, which can be a handicap for large-scale epi-

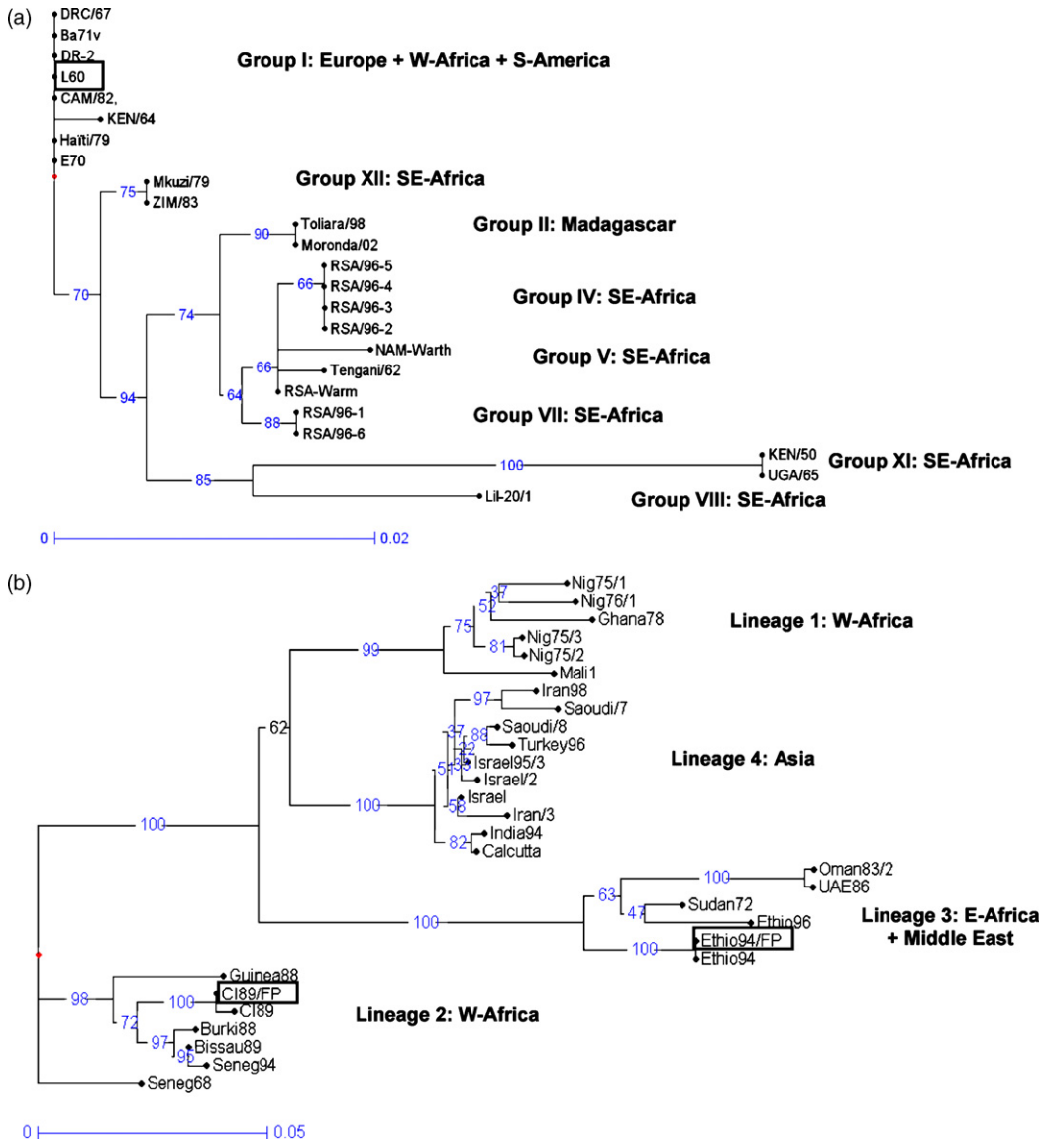


Fig. 2. Phylogenetic analysis of ASFV (a) and PPRV (b). Sequences derived from blood-dried filter papers collected on infected pigs or goats are shown in boxes. Sequencing and analysis of sequences were done as indicated in Material and methods. Consensus trees were generated on 1000 replicates. Only bootstraps higher than 60 are shown. The strains L60 (a), CI89 and Ethio94 (b) amplified and sequenced from filter papers were found in Group I Lineages 2 and 3, respectively as expected according to the work of Bastos et al. (2003), Dhar et al. (2002) and Kwiatek et al. (2007).

demiological surveys. Also, the elution step before PCR adds 1 h to the protocol. Although FTA cards may improve the sensitivity of detection, several factors may make them inconvenient and expensive for routine large-scale surveys. Our data show that Whatman 3MM papers are an effective cheaper alternative. Subsequently, a satisfactory analytical sensitivity was obtained with these filter papers (around 1–2 $TCID_{50}$), which is considered as sensitive enough to detect ASFV infected pigs as illustrated in the *in vivo* trials. In this study, this approach was validated for a DNA virus known to be highly resistant in the environment and a RNA virus which is considered to be very labile outside the host. For the two viruses used in this study, the defined protocol gave an excellent sensitivity (around 1–2 $TCID_{50}$ detected). The discrepancy between the number of DNA copies and tissue culture infectious doses 50 detected (1–10 for ASFV and PPRV) can be ascribed to the fact that viruses which are non-viable but still contain undamaged genome or free genomes are probably released during *in vitro* replication as already reported *in vitro* but also *in vivo* in other virus infection systems (MacLachlan et al., 1994; Tedder et al., 1998; Spagnuolo-Weaver et al., 1998). With PPRV, an important difference in the number of DNA and RNA copies detected on filter papers was noticed: 1 DNA copy versus 2.10^4 RNA copies. This difference is probably more resulting from a reduced yield of cDNA during the reverse transcription of RNA than from RNA denaturation on filter papers. Indeed, Katz et al. (2002) also found a sensitivity of 10^4 copies of measles virus RNA in a single-step RT-PCR done on soluble RNA. Previous studies have shown that filter papers enabled storage of double or single stranded RNA viruses for 1 month at 37 °C (Pitcovski et al., 1999; De Swart et al., 2001) and at least 3 months at 30 °C for ASFV (Guy-Gonzague et al., 2003). In this study, filter papers were allowed to dry shortly after soaking, once dried,

they were stored for extended periods of time. It is shown that the stability can be longer than 9 months at 37 °C for ASFV, thus reinforcing the excellent capacity of filter papers containing dried blood to store genetic material. With the single stranded RNA virus used as a model in this study, filter papers containing dried blood could maintain nucleic acids more than 3 months at 32 °C. PPRV belongs to the same genus *Morbillivirus* as measles virus. In a previous study, measles virus RNA could be stored for 1 month at 37 °C and the sensitivity of the detection was 100 $TCID_{50}$ for a single PCR and 3 $TCID_{50}$ for a nested PCR (Katz et al., 2002). This work shows that with another *morbillivirus* an analytical sensitivity of 1–2 $TCID_{50}$ can be achieved with a single PCR. Although, a relatively limited number of paired samples (blood from EDTA tubes and from filter papers) was used, a satisfactory agreement was found between the two collection materials. Discrepant results were seen for two couples of filter paper/EDTA blood out of 19. These couples pertaining to the trial 2 and collected at days 3 and 7 post-challenge provided inverse results −/+ and +/−, thus giving no clear advantage to one of these collection materials. The use of filter papers is also compatible with genetic characterization of the strains as shown in this study and others (Nerurka et al., 1993; Pitcovski et al., 1999; De Swart et al., 2001). Direct sequencing from filter papers allowed to rapidly group the strains into phylogeographic dendogrammes. In that case, molecular sequencing of the strains is particularly useful to trace the geographic origin of the infection (Mosquera et al., 2004; Verbeeck et al., 2006).

5. Conclusion

The new protocol proposed in this study is rapid, does not need previous extraction of nucleic acids, limits the risk of cross-contamination between samples, allows long term-storage of blood at relatively high temperatures and simplifies shipment to the laboratory without the need for cold chain. The unique constraint is the necessity of using a set of forceps and scissors for individual preparation of filter papers in the direct PCR. However, these instruments can be easily decontaminated and used again. The method described here could be adapted to any DNA and RNA viruses using peripheral blood for circulation within the host. Development perspectives for this method are the use of the same filter paper samples for the combined direct detection of nucleic acids, antibodies and antigens as already done for measles virus (De Swart et al., 2001). In addition, the use of filter papers for quantitative detection of virus genomes will be shortly evaluated as already done before for the duck hepatitis B virus (Wang et al., 2002).

Acknowledgements

This study was partially granted by European Union grant no QLRT-2000-02216 (5th Framework, Quality of life), EU Pan-African programme for the Control of Epizootics (PACE, REG/5005/005) and EPIZONE network of Excellence noFOOD-CT-2006-016236. African swine fever virus strains from Madagascar were kindly provided by the Direction de la Santé Animale et du Phytosanitaire du Ministère de l'Agriculture de l'Elevage et de la Pêche of Madagascar.

References

Abe, K., Konomi, N., 1998. Hepatitis C virus RNA in dried serum spotted onto filter paper is stable at room temperature. J. Clin. Microbiol. 36, 3070–3072.

Basto, A.P., Portugal, R.S., Nix, R.J., Cartaxeiro, C., Boinas, F., Dixon, L.K., Leitao, A., Martins, C., 2006. Development of a nested PCR and its internal control for the detection of African swine fever virus (ASFV) in Ornithodoros erraticus. Arch. Virol. 151, 819–826.

Bastos, A.D., Penrith, M.L., Cruciere, C., Edrich, J.L., Hutchings, G., Roger, F., Couacy-Hymann, E.R., Thomson, G., 2003. Genotyping field strains of African swine fever virus by partial p72 gene characterisation. Arch. Virol. 148, 693–706.

Beck, I.A., Drennan, K.D., Melvin, A.J., Mohan, K.M., Herz, A.M., Alarcon, J., Piscoya, J., Velazquez, C., Frenkel, L.M., 2001. Simple, sensitive, and specific detection of human immunodeficiency virus type 1 subtype B DNA in dried blood samples for diagnosis in infants in the field. J. Clin. Microbiol. 39, 29–33.

Behets, F., Kashamuka, M., Pappaioanou, M., Green, T.A., Ryder, R.W., Batter, V., George, J.R., Hannon, W.H., Quinn, T.C., 1992. Stability of human immunodeficiency virus type 1 antibodies in whole blood dried on filter paper and stored under various tropical conditions in Kinshasa. Zaire. J. Clin. Microbiol. 30, 1179–1182.

Chaisomchit, S., Wichajarn, R., Janejai, N., Chareonsiriwatana, W., 2005. Stability of genomic DNA in dried blood spots stored on filter paper. Southe. Asian J. Trop. Med. Public Health 36, 270–273.

Couacy-Hymann, E., Roger, F., Hurard, C., Guillou, J.P., Libeau, G., Diallo, A., 2002. Rapid and sensitive detection of peste des petits ruminants virus by a polymerase chain reaction assay. J. Virol. Methods 100, 17–25.

Couacy-Hymann, E., Bodjo, C., Danho, T., Libeau, G., Diallo, A., 2005. Evaluation of the virulence of some strains of peste-des-petits-ruminants virus (PPRV) in experimentally infected West African dwarf goats. Vet. J. 23, 178–183.

Dhar, P., Sreenivasa, B.P., Barrett, T., Corteyn, M., Singh, R.P., Bandyopadhyay, S.K., 2002. Recent epidemiology of peste des petits ruminants virus (PPRV). Vet. Microbiol. 88, 153–159.

De la C Herrera, R., Cabrera, M.V., Garcia, S., Gilart, M., 2006. IgM antibodies to dengue virus in dried blood on filter paper. Clin. Chim. Acta 367, 204–206.

De Swart, R.L., Nur, Y., Abdallah, A., Kruining, H., El Mubarak, H.S., Ibrahim, S.A., Van Den Hoogen, B., Groen, J., Osterhaus, A.D., 2001. Combination of reverse transcriptase PCR analysis and immunoglobulin M detection on filter paper blood samples allows diagnostic and epidemiological studies of measles. J. Clin. Microbiol. 39, 270–273.

Dubay, S.A., Rosenstock, S.S., Stallknecht, D.E., Devos Jr., J.C., 2006. Determining prevalence of bluetongue and epizootic hemorrhagic disease viruses in mule deer in Arizona (USA) using whole blood dried on paper strips compared to serum analyses. J. Wild L. Dis. 42, 159–163.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14, 685–695.

Guy-Gonzague, M., Roger, F., Rousset, D., Randriamparany, T., Cruciere, C., 2003. Int. J. Appl. Res. Vet. Sci. 1 (2), http://www.jarvm.com/articles/Vol1Iss2/Gonzague.htm.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: Nucleic Acids Symposium Series 41, pp. 95–98.

Hattermann, K., Soike, D., Grund, C., Mankertz, A., 2002. A method to diagnose Pigeon circovirus infection *in vivo*. J. Virol. Methods 104, 55–58.

Hogrefe, W.R., Ernst, C., Su, X., 2002. Efficiency of reconstitution of immunoglobulin g from blood specimens dried on filter paper and utility in herpes simplex virus type-specific serology screening. Clin. Diagn. Lab. Immunol. 9, 1338–1342.

Hutet, E., Chevallier, S., Eloit, M., Touratier, A., Blanquefort, P., Albina, E., 2003. Porcine reproductive and respiratory syndrome antibody detection on filter discs. Rev. Sci. Tech. 22, 1077–1085.

Jakobsson, U., Westergren, A., 2005. Statistical methods for assessing agreement for ordinal data. Scand. J. Caring Sci. 19, 427–431.

Kaerber, G., 1931. Beitrag zur kollektiven behandlung pharmakologischer. Reihenversuche, Naunyn Schmiedebergs. Arch. Pharmakol. Exp. Pathol. 142, 480–483.

Kailash, U., Hedau, S., Gopalkrishna, V., Katiyar, S., Das, B.C., 2002. A simple 'paper smear' method for dry collection, transport and storage of cervical cytological specimens for rapid screening of HPV infection by PCR. J. Med. Microbiol. 51, 606–610.

Katz, R.S., Premenko-Lanier, M., McChesney, M.B., Rota, P.A., Bellini, W.J., 2002. Detection of measles virus RNA in whole blood stored on filter paper. J. Med. Virol. 67, 596–602.

Kwiatek, O., Minet, C., Grillet, C., Hurard, C., Carlsson, E., Karimov, B., Albina, E., Diallo, A., Libeau, G., 2007. Peste des petits ruminants (PPR) outbreak in Tajikistan. J. Comp. Path. 136, 111–119.

Leitao, A., Cartaxeiro, C., Coelho, R., Cruz, B., Parkhouse, R.M., Portugal, F., Vigario, J.D., Martins, C.L., 2001. The non-haemadsorbing African swine fever virus isolate ASFV/NH/P68 provides a model for defining the protective anti-virus immune response. J. Gen. Virol. 82, 513–523.

Li, C.C., Beck, I.A., Seidel, K.D., Frenkel, L.M., 2004. Persistence of Human immunodeficiency virus type 1 subtype B DNA in dried-blood samples on FTA filter paper. J. Clin. Microbiol. 42, 3847–3849.

Lopez-Otin, C., Freije, J.M., Parra, F., Mendez, E., Vinuela, E., 1990. Mapping and sequence of the gene coding for protein p72, the major capsid protein of African swine fever virus. Virology 175, 477–484.

MacLachlan, N.J., Nunamaker, R.A., Katz, J.B., Sawyer, M.M., Akita, G.Y., Osburn, B.I., Tabachnick, W.J., 1994. Detection of bluetongue virus in the blood of inoculated calves: comparison of virus isolation, PCR assay, and *in vitro* feeding of Culicoides variipennis. Arch. Virol. 136, 1–8.

Moscoso, H., Raybon, E.O., Thayer, S.G., Hofacre, C.L., 2005. Molecular detection and stereotyping of infectious bronchitis virus from FTA filter paper. Avian Dis. 49, 24–29.

Mosquera, M.M., Echevarria, J.E., Puente, S., Lahulla, F., de Ory, F., 2004. Use of whole blood dried on filter paper for detection and genotyping of measles virus. J. Virol. Methods 117, 97–99.

Natarajan, P., Trinh, T., Mertz, L., Goldsborough, M., Fox, D.K., 2000. Paper-based archiving of mammalian and plant samples for RNA analysis. Biotechniques 29, 1328–1333.

Nerurka, V.R., Babu, P.G., Song, K.J., Melland, R.R., Gnanamuthu, C., Saraswathi, N.K., Chandy, M., Godec, M.S., John, T.J., Yanagihara, R., 1993. Sequence analysis of Human T cell lymphotropic virus type 1 strains from southern India: gene amplification and direct sequencing from whole blood blotted onto filter paper. J. Gen. Virol. 74, 2799–2805.

Perrier, X., Flori, A., Bonnet, F., 2003. Data analysis methods. In: Hamos, P., Seguin, M., Perrier, X., Glaszmann, J.C. (Eds.), Genetic Diversity of Cultivated Tropical Plants. Gifield Science Publishers, Montpellier, pp. 43–76.

Pitcovski, J., Shmueli, E., Krispel, S., Levi, N., 1999. Storage of viruses on filter paper for genetic analysis. J. Virol. Methods 83 (1–2), 21–26.

Prado, I., Rosario, D., Bernardo, L., Alvarez, M., Rodriguez, R., Vazquez, S., Guzman, M.G., 2005. PCR detection of dengue virus using dried whole blood spotted on filter paper. J. Virol. Methods 1, 75–81.

Roeder, P.L., Abraham, G., Kenfe, G., Barrett, T., 1994. Peste des petits ruminants in Ethiopian goats. Trop. Anim. Health Prod. 26, 69–73.

Saitou, N., Nei, M., 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sierra, M.A., Carrasco, L., Gomez-Villamandos, J.C., Martin de las Mulas, J., Mendez, A., Jover, A., 1990. Pulmonary intravascular macrophages in lungs of pigs inoculated with African swine fever virus of differing virulence. J. Comp. Pathol. 102, 323–334.

Spagnuolo-Weaver, M., Walker, I.W., McNeilly, F., Calvert, V., Graham, D., Burns, K., Adair, B.M., Allan, G.M., 1998. The reverse transcription polymerase chain reaction for the diagnosis of porcine reproductive and respiratory syndrome: comparison with virus isolation and serology. Vet. Microbiol. 62, 207–215.

Tedder, R.S., Kaye, S., Loveday, C., Weller, I.V., Jeffries, D., Norman, J., Weber, J., Bourelly, M., Foxall, R., Babiker, A., Darbyshire, J.H., 1998. Comparison of culture- and non-culture-based methods for quantification of viral load and resistance to antiretroviral drugs in patients given zidovudine monotherapy. J. Clin. Microbiol. 36, 1056–1063.

Van de Peer, Y., De Wachter, R., 1993. TREECON: a software package for the construction and drawing of evolutionary trees. Comput. Appl. Biosci. 9, 177–182.

Verbeeck, J., Maes, P., Lemey, P., Pybus, O.G., Wollants, E., Song, E., Nevens, F., Fevery, J., Delport, W., Van der Merwe, S., Van Ranst, M., 2006. Investigating the origin and spread of hepatitis C virus genotype 5a. J. Virol. 80, 4220–4226.

Wang, C.Y., Giambrone, J.J., Smith, B.F., 2002. Detection of duck hepatitis B virus DNA on filter paper by PCR and SYBR green dye-based quantitative PCR. J. Clin. Microbiol. 40, 2584–2590.

Yourno, J., Conroy, J., 1992. A novel polymerase chain reaction method for detection of human immunodeficiency virus in dried blood spots on filter paper. J. Clin. Microbiol. 30, 2887–2892.

Yozawa, T., Kutish, G.F., Afonso, C.L., Lu, Z., Rock, D.L., 1994. Two novel multigene families, 530 and 300, in the terminal variable regions of African swine fever virus genome. Virology 202, 997–1002.

Yu, M., Morrissy, C.J., Westbury, H.A., 1996. Strong sequence conservation of African swine fever virus p72 protein provides the molecular basis for its antigenic stability. Arch. Virol. 141, 1795–1802.

Zhou, H., Hickford, J.G., Fang, Q., 2006. A two-step procedure for extracting genomic DNA from dried blood spots on filter paper for polymerase chain reaction amplification. Anal. Biochem. 354, 159–161.

Zollner, B., 2004. Surveillance of the molecular epidemiology of hepatitis B virus in industrialized countries: necessary despite low prevalence and an available, effective vaccine? Clin. Infect. Dis. 39, 953–954.

Le protocole que nous avons proposé au cours de cette étude est rapide, il ne nécessite pas l'extraction des acides nucléiques et limite les risques de contaminations croisées entre les échantillons. En plus de permettre le transport et le stockage des échantillons en s'affranchissant de la chaine du froid, le papier filtre permet une longue conservation des échantillons de sang à des températures relativement élevées. L'unique contrainte liée à son utilisation est de disposer de pinces et de ciseaux pour préparer les fragments utilisés pour effectuer la PCR directe. Mais ces instruments sont réutilisables puisque facilement décontaminés. Cette méthode est adaptable à tout type de virus à ARN ou à ADN génomique générant une virémie chez l'hôte. Pour étendre la gamme des tests diagnostics réalisables avec cette méthode, la détection des anticorps et des antigènes va être mise en œuvre, en nous basant sur ce qui a été développé pour le virus de la rougeole (De Swart *et al.* 2001). De même, comme il a été fait pour avec le virus de l'hépatite B du canard (Wang *et al.* 2002), la PCR directe en temps réel à partir de papier filtre est en cours de développement.

Cette approche permettant la collecte et le transport des échantillons a été largement utilisée, depuis la parution de cet article, dans plusieurs pays africains, notamment à Madagascar, en Côte d'Ivoire ainsi qu'en RDC (communications personnelles). La double possibilité offerte par cette méthode de détecter puis de caractériser les souches circulantes de virus PPA lui confère un avantage déterminant pour la surveillance épidémiologique et donc pour le contrôle de la maladie, mais aussi pour l'obtention de séquences supplémentaires pour analyser l'évolution du virus PPA.

Par ailleurs, la même méthode vient d'être récemment validée pour la détection d'anticorps sériques, la détection moléculaire du génome par PCR en temps réel directement sur papier buvard et l'isolement viral à partir de buvard, témoignant du grand intérêt de la méthode pour une utilisation en surveillance dans les zones tropicales (article en préparation).

Partie 2

Des reconstructions phylogénétiques à la caractérisation de filiation entre isolats.

La caractérisation des souches circulantes, tant au sein de la faune sauvage, qu'au sein des élevages porcins est essentielle en termes de compréhension de l'épidémiologie de la maladie. Ainsi, le traçage des souches permet-il de mieux comprendre les voies empruntées par le virus pour coloniser de nouveaux territoires, ou la façon dont il parvient à se maintenir dans une région donnée. Cette compréhension peut alors avoir une traduction en termes de mesures sanitaires de biosureté, aidant de fait à l'endiguement de la propagation ou de la résurgence de la maladie. Lorsque la PPA est entrée en 1998 à Madagascar, nous nous sommes attachés à échantillonner régulièrement les virus circulants sur l'île et à les caractériser sur le plan moléculaire. Nous avons effectué des reconstructions moléculaires sur la base de méthodes modernes probabilistes (maximum de vraisemblance et méthode bayésienne) pour tenter de discriminer géographiquement et chronologiquement les isolats circulants. Si cette discrimination a été peu probante dans un premier temps, elle nous a revanche permis rapidement de contribuer à caractériser la sortie du virus depuis Madagascar ou de son berceau initial en Afrique du sud-est (Mozambique) vers le Caucase en 2007. Une étude phylogénétique visant à caractériser la souche responsable de l'épidémie qui sévit désormais en Europe a été menée et a en effet permis de révéler que le virus PPA en cause appartenait au même génotype que le virus Madagascar.

Ces travaux ont été valorisés par un second article présenté dans ce qui suit.

African Swine Fever Virus Isolate, Georgia, 2007

Rebecca J. Rowlands, Vincent Michaud, Livio Heath, Geoff Hutchings, Chris Oura, Wilna Vosloo, Rahana Dwarka, Tinatin Onashvili, Emmanuel Albina, and Linda K. Dixon

African swine fever (ASF) is widespread in Africa but is rarely introduced to other continents. In June 2007, ASF was confirmed in the Caucasus region of Georgia, and it has since spread to neighboring countries. DNA fragments amplified from the genome of the isolates from domestic pigs in Georgia in 2007 were sequenced and compared with other ASF virus (ASFV) isolates to establish the genotype of the virus. Sequences were obtained from 4 genome regions, including part of the gene *B646L* that encodes the p72 capsid protein, the complete *E183L* and *CP204L* genes, which encode the p54 and p30 proteins and the variable region of the *B602L* gene. Analysis of these sequences indicated that the Georgia 2007 isolate is closely related to isolates belonging to genotype II, which is circulating in Mozambique, Madagascar, and Zambia. One possibility for the spread of disease to Georgia is that pigs were fed ASFV-contaminated pork brought in on ships and, subsequently, the disease was disseminated throughout the region.

African swine fever (ASF), classified as a notifiable disease by the World Organisation for Animal Health (OIE), causes an acute hemorrhagic fever in domestic pigs. It often results in major economic losses because of the high rates of illness and death associated with the disease. ASF has the potential to spread rapidly and since a vaccine is currently not available, control options are limited to rapid diagnosis of the disease and culling of infected animals and animals in contact with them.

ASF virus (ASFV) infects wildlife hosts and ticks of *Ornithodoros* spp., and these can provide a reservoir of virus that is not possible to eliminate. In Africa, where ASF is widespread, the virus causes long-term, persistent infections—but no clinical manifestation of disease—in warthogs (*Phacochoerus africanus*) and bushpigs (*Potamochoerus porcus*) (*1*). In contrast, ASFV causes clinical disease and high numbers of deaths in wild boars (*Sus scrofa*). The disease has been reported in pigs from most sub-Saharan countries and continues to spread to previously uninfected countries within the region. In 1998, ASF was reported in Madagascar for the first time; it is now considered to be endemic. At the end of 2007, ASF was introduced on a second Indian Ocean island, Mauritius (*2*).

Once introduced into countries, ASF is difficult to eradicate for several reasons, including the presence of wildlife reservoirs, lack of a vaccine, insufficient laboratory support for rapid and accurate diagnosis, and inadequate funding for veterinary services to enforce the appropriate control measures. This situation was amply demonstrated in Portugal and Spain, where the disease remained endemic until the 1990s, after its introduction into Portugal in 1957 and again in 1960. Other European countries, the Caribbean, and Brazil have had outbreaks of ASF, but extensive control programs have led to successful eradication, with the exception of Sardinia, where ASF has remained endemic since 1982 (*1*).

The genetic characterization of the viral strain associated with disease outbreaks is important for tracing possible sources of infection and ensuring that appropriate diagnostic reagents are used. ASFV isolates have previously been characterized by restriction enzyme site mapping or sequencing of different genome regions. Partial sequencing

Author affiliations: Institute for Animal Health, Pirbright, UK (R.J. Rowlands; G. Hutchings, C. Oura, L.K. Dixon); Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Montpellier, France (V. Michaud, E. Albina); Agricultural Research Council–Onderstepoort Veterinary Institute, Onderstepoort, South Africa (L. Heath, W. Vosloo, R. Dwarka); University of Pretoria, Pretoria, South Africa (W. Vosloo); and Laboratory of Ministry of Agriculture of Georgia, Tbilisi, Georgia (T. Onashvili)

DOI: 10.3201/eid1412.080591

of the *B646L* gene encoding the major capsid protein p72 has so far led to the identification of 22 ASFV genotypes. Twenty-one of these genotypes were identified in isolates from domestic pigs or from wildlife hosts in eastern and southern Africa. The level of diversity between isolates from these regions is attributed to the long-term evolution of virus within wildlife hosts. In contrast to the other genotypes, genotype I predominantly comprises isolates from domestic pigs in West and Central Africa, Europe, the Caribbean, and Brazil obtained during a 40-year period since 1957. Isolates belonging to genotype I share considerably higher sequence identity across the p72 gene compared to isolates from the sylvatic cycle, which suggests that this genotype probably evolved from a single source introduction (*3*–*5*). PCR amplification and sequencing of more variable genome regions have been used to distinguish between closely related isolates and identify virus subgroups within several of the 22 genotypes (*4*). The additional genome regions that have been described thus far include the *E183L* and *CP204L* gene regions, encoding the p54 and p30 proteins, respectively, as well as the central variable region within the open reading frame (ORF) *B602L*.

ASFV can infect pigs by a variety of mechanisms, including direct contact between pigs, bites from infected ticks, indirect transmission by means of fomites, and ingestion of infected meat. The main route by which ASF infections spread over long distances is thought to be infected meat products. The transcontinental spread of ASF has been a relatively rare event, and it was unexpected when, in June 2007, cases of ASF affecting domestic pigs in the Caucasus region of the former Soviet republic of Georgia were confirmed by the OIE ASF reference laboratory (*6*). It has been suggested that the outbreak started in April 2007 near the Black Sea Port of Poti. Catering waste, including infected pig meat from ships in the port, is considered to be the most likely source of the infection. As of July 9, 2007, the outbreak had spread to 56 of 61 districts in Georgia. Reports to OIE indicated that >80,000 pigs had died or been destroyed in Georgia. Outbreaks of ASF were also reported in neighboring regions, including the autonomous republic of Abkhazia (*7*). On August 29, 2007, ASF was confirmed in Armenia and on November 4, 2007 (*8*), in Nagorno-Karabakh (*9*), a de facto independent republic that is officially part of Azerbaijan and near its border with Armenia. On November 5, 2007, infection of a wild boar was confirmed in the Russian Republic of Chechnya near the border with Georgia. To control the spread of disease, wild boars were killed in 17 different regions in Chechnya, and the slaughter of the entire pig population was ordered (*10*). Further outbreaks of ASF were reported in Nagorno-Karabakh in April 2008, where it is believed that ≈8,500 pigs have died as a result of disease since the beginning of the outbreaks.

Here we describe the genetic characterization of the ASFV isolates implicated in the 2007 outbreak of the disease in Georgia. Results of the analysis showed that the Georgia isolates group within genotype II, which suggests that the virus is closely related to ASFV isolates typically found in Mozambique, Madagascar, and Zambia (*4*,*11*,*12*).

Materials and Methods

Virus Isolates

In June 2007, samples were collected from 2 pigs that were showing clinical signs of ASF. The first pig originated from the Imereti Province in western Georgia; the second was sampled in the Kakheti Province in eastern Georgia. Five tissues samples were collected from each pig, including serum and samples from the kidney, spleen, lung, and lymph nodes. The samples were subsequently submitted to the OIE reference laboratory, Institute for Animal Health, Pirbright, United Kingdom. The presence of ASFV in these samples was confirmed by pathogen isolation on primary leukocyte cultures, real-time PCR, and ELISA (*13*,*14*).

Viral DNA Extraction, PCR Amplification, and Sequencing

Viral DNA was extracted directly from cell culture isolates or from suspensions of clinical samples by using the High Pure Viral Nucleic Acid Kit (Roche, Indianapolis, IN, USA) following the manufacturer's guidelines. The extracted DNA was used as template for the amplification of the respective gene regions. Details of isolates studied are shown in the online Appendix Table (available from www.cdc.gov/EID/content/14/12/1870-appT.htm).

PCRs were performed with the Accuprime *Pfx* DNA polymerase (Invitrogen, Carlsbad, CA, USA). Reactions contained 22.5 μL Accuprime *Pfx* Supermix, 100 ng DNA, and a final concentration of 200 nmol/L of each primer in a total reaction volume of 25 μL. Thermocycling condition included a 2-min denaturation step of 95°C, followed by 35 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 68°C with a 10-min elongation step at 68°C. Part of the gene encoding the p72 gene was amplified by using the primers P72-D and P72-U (*3*), which amplify a 478-bp fragment from the 3′ end of the *B646L* gene. The primer pair ORF9L-F (5′-AATGCGCTCAGGATCTGTTAAATCGG-3′) and ORF9L-R (5′-TCTTCATGCTCAAAGTGCGTATACCT-3′) was used to amplify a region from the central variable genome within the ORF B602L (*16*); E183L-F (5′-TCACCGAAGTGCATGTAATAAACG-3′) and E183L-R (5′-TCTGTAATTTCATTGCGGCCACAACATT-3′) were used to amplify a 681-bp fragment of the *E183L* gene. Primer pairs p30-F (5′-ATGAAAATGG AGGTCATCTTCAAAAC-3′) and p30-R (5-AAGTT

TAATGACCATGAGTCTTACC-3′) were used to amplify 521 bp of the *CP204L* gene.

Primers used for the amplification of p72, p54, p30, and B602L gene regions, as described above, were used in the respective sequencing reactions. Sequencing of PCR products was performed by using the Dye Terminator Cycle Sequencing Quick Start Kit (Beckman Coulter, Fullerton, CA, USA). Thermocycling consisted of 30 cycles of 96°C for 20 s, 50°C for 20 s, and 60°C for 3 min. Completed reactions were processed following the manufacturer's instructions. Data was processed by using the default sequence analysis parameters and analyzed with Beckman Coulter CEQ 8000 software.

Sequence Analysis

Analysis of sequence data was performed with Beckman Coulter CEQ8000 software, Chromas (www.technelysium.com.au), BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html), and ClustalX version 1.83 (www.clustal.org). A summary of the sequences is shown in the online Appendix Table.

Phylogenetic analysis was conducted by means of the "criterion of neighborhood based on the principle of parsimony" (www.megasoftware.net/index.html; *17*,*18*), selecting the correction of Kimura (*19*). Bootstrap confidence values were calculated on 1,000 replicates according to the maximum likelihood approach of Felsenstein (*20*).

Results

Partial Sequence of *B646L* Gene Encoding the p72 Capsid Protein

Sequence analysis of the *B646L* gene has been used extensively for phylogenetic analysis of ASFV isolates (*3*,*5*,*15*) by focusing on a 478-bp fragment corresponding to the C-terminal end of the *B646L* gene that broadly defines the virus genotypes. Twenty-two genotypes (*4*) have thus far been identified by analyzing this region of the viral genome.

The *B646L* partial sequences from each of the 5 tissue samples from the east and west Georgian samples showed that they were identical at the nucleotide level (results not shown). Comparison of these sequences to other isolates of known genotypes identified the Georgia 2007 sequence as falling within *B646L* genotype II (Figure), together with 1 isolate from Zambia (Lus 1/93), isolated from a domestic pig after an outbreak of ASF in 1991 (*10*); 9 from Mozambique (Moz 60–98, Moz 61–98, Moz 63–98, Moz 70–98, Moz 77–98, Moz 1/02, Moz 2/02, Moz 1/03, and Moz 1/05), obtained from outbreaks in 1998–2005 (*5*,*11*,*12*); and a pig isolate from Madagascar (Mad 1/98), obtained after the first introduction in 1998 (*3*,*21*).

Sequence Analysis of *B602L* Region

The central variable region of the ORF *B602L* is characterized by tetrameric repeats, the number and composition which can be used to distinguish between closely related isolates (*16*). Sequence analysis of this region from the *B602L* gene (also designated central variable region ORF9L, 9RL) of >100 ASFV isolates has shown that the number of tandem repeat tetramers in individual genomes may vary from 7 to 34. Twenty-two sequence variants of the 4-aa repeats have also been identified (*15*).

Amplification of the *B602L* variable fragment from each of the east and western Georgian isolates yielded PCR products of ≈200 bp, which corresponded in size and sequence to the other genotype II isolates with 10-aa tetramers. The sequences of this region differed from that of all other genotypes (online Appendix Figure 1, available from www.cdc.gov/EID/content/14/12/1870-appF1.htm). Despite also containing 10 copies of amino acid tetramers, the *B602L* sequence of 2 South African isolates from genotype XXI differed from Georgia 2007 and the other genotype II isolates.

Sequence Analysis of *E183L* Gene Encoding Protein p54

Amplification of the fragment containing the complete *E183L* gene from all the Georgian isolates produced PCR products of ≈550 bp, which were identical in sequence

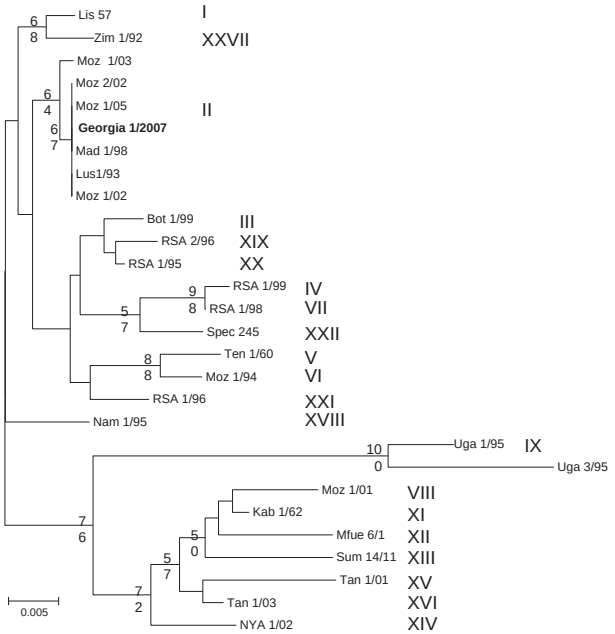


Figure. Phylogram depicting the *B646L* gene relationships of selected isolates representative of the 22 African swine fever virus genotypes. Because all the Georgian isolates had identical nucleotide sequences, only 1 isolate is presented in the tree (in **boldface**). The consensus tree was generated from 1,000 replicates; only bootstraps >50% are shown. Genotypes are indicated in roman numerals. Moz, Mozambique. Scale bar indicates number of nucleotide substitutions per site.

(results not shown). Amplification of this fragment from other isolates from Africa, Europe, and Madagascar produced fragments that ranged from 528 bp to 600 bp. The discrepancy in the size of the respective fragments is due to sequences encoding 2 arrays of amino acid repeats, which vary in number and sequence. The nucleotide sequence of the *E183L* gene from the Georgia 2007 isolate was identical to sequences from 5 Madagascar isolates obtained from outbreaks that in 1998–2003 (Mad 1/98, Ampani/99, Tolagna/99, Chrome/01, and Antani/03) and 2 Mozambican isolates (Moz 1/02, Moz 2/02) (online Appendix Figure 2, available from www.cdc.gov/EID/content/14/12/1870-appF2.htm). However, the p54 nt and protein sequences of isolates Moz 1/03, Moz 1/05, and Lus 1/93 differed from that of the Georgia 2007 isolates; the latter contained a single deletion between positions 341 and 355, resulting in a 5-aa deletion within the central portion of the protein. The nucleotide sequence of the 2 other isolates from Mozambique obtained in 2003 and 2005 (Moz 1/03 and Moz 1/05) were identical to each other but differed from that of the Georgia 2007 isolates at several positions throughout the gene region (online Appendix Figure 2).

Sequence Analysis of *CP204L* Gene Encoding p30 Protein

Amplification of a fragment containing the *CP204L* gene from each of 2 Georgian isolates produced a PCR product of ≈550 bp. As was the case in all the other gene regions, the sequence of the 2 Georgian isolates was identical across the length of the gene (results not shown). The nucleotide sequences of the Georgia isolates were unique within genotype II but shared a high degree of similarity with the isolates from Madagascar, Zambia, and Moz 2/02 (nucleotide identity >99%). In contrast, isolates implicated in the most recent outbreaks of the disease in Mozambique, isolates Moz 1/03 and Moz 1/05, differed from the Georgian isolates by >2.5% at the nucleotide level (online Appendix Figure 3, available from www.cdc.gov/EID/content/14/12/1870-appF3.htm).

Discussion

We analyzed the sequence of 4 genomic fragments of the ASFV genome to characterize the viruses responsible for the outbreak of ASF in Georgia in 2007. The 4 regions of the genome—*B646L, E183L, CP204L,* and the variable region within the ORF *B602L*—were amplified by PCR. The nucleotide and amino acid sequences of these ORFs from samples collected at 2 different geographic locations in Georgia were then compared with ASFV isolates from other regions of the world. Because all DNA and amino acid sequences for each genome region from all tissue samples obtained from Georgia that we tested were identical, we concluded that the ASF outbreaks in Georgia and the surrounding regions were probably due to a single introduction of the virus. Sequence analysis of the p72 gene region placed the Georgian isolate within genotype II together with isolates from Madagascar, Mozambique, and Zambia (*3*–*5*). Genotype II occurred in Mozambique in outbreaks in 1998–2005 (*12*) and affected the northeastern provinces of Cabo Delgado and Nampula (which were most recently affected in 2004), the northwestern province of Tete, and the southern province of Maputo (most recently in 2005) (*12*). Three other genotypes of ASFV have also been identified as having occurred in Mozambique—genotypes II, V, and VI (*12*).

The genotype II Madagascar isolate, MAD 1/98, was obtained from a domestic pig in 1998 during the first outbreak of ASF that affected the island country. The more recent Madagascar pig isolates obtained in 1999–2003 are presumed to have derived from this first introduction because they belong to the same genotype. Mozambique has been speculated to be the most likely source of infection for the 1998 ASFV outbreaks occurring in Madagascar because the isolates from Mozambique were genotype II and identical across the *B602L* region (*22*). Before 1998, the island of Madagascar was free of the disease (*21*,*22*). The genotype II isolate from Zambia (Lus 1/93) was isolated from an infected domestic pig in 1991, whereas the viruses from Mozambique were isolated from domestic pigs during outbreaks of the disease along the eastern coast of the country in 2002–2005.

Further sequence analysis was performed on 2 other conserved regions of the ASFV genome; the ORFs *E183L* and *CP204L*, which encode the structural proteins p54 and p30, respectively. Analysis of the *E183L* gene showed that the Georgia 2007 isolates were most closely related to 4 isolates from Madagascar, which were in circulation in 1999–2003, and 2 isolates from Mozambique but distinguishable from the group II isolate Lus 1/93 and the Mozambique isolates Moz 1/03 and Moz 1/05. Similarly, analysis of the *CP204L* gene encoding p30 showed the Georgia 2007 isolates were distinguishable from all other isolates, although they were most closely related to the 4 isolates from Madagascar in circulation in 1999–2003, the Zambian Lus 1/93 isolate, and one of the Mozambique isolates (2/02).

Fragment size analysis has identified *B602L* as the most variable genome region (*15*). The variable region of *B602L* contains amino acid tetramers that vary in number and type. Sequence analysis of the *B602L* gene from the Georgian isolates identified 5 different amino acid tetramer sequences encoded in this genome region. One of these tetramer sequences was CTST, which is one of the less common tetramer sequences (*15*). The sequence of the *B602L* variable region from the Georgian isolate grouped it with isolates in circulation in Madagascar (1999–2003) as well as isolates from Mozambique from outbreaks in 1960, 1961, 1963, 1970, and 1998.

The first case of ASF in Georgia was observed in the Samegrelo region on the west coast, which suggests a possible connection to the port of Poti on the Black Sea. One possibility is that that the virus entered Georgia through meat products since ASFV may remain viable for long periods in infected pig tissues, meat, and processed pig products. Most pigs in Georgia are kept on a free-ranging, scavenging system, and so access to or swill feeding of dumped port waste is possible. However, several events would be required to cause an outbreak, making this a relatively rare event and providing an explanation for the relatively few incidents of transcontinental spread of ASFV. Our analysis showed that the Georgia strain is most similar to isolates from Madagascar. However, since few ASFV samples are submitted for genotyping, it is possible that viruses belonging to genotype II may be more widespread. However, it seems likely that the source of infection of the Georgia 2007 outbreak is from the eastern side of southern Africa or Madagascar rather than west or central Africa or Sardinia.

Acknowledgments

We thank François Roger for arranging access to the Madagascar isolates.

Part of the work on African isolates was supported by the Wellcome Trust project "African swine fever virus: Development of vaccines and epidemiological investigations" (application 075813, 2005–2010). Part of the work was funded by the European Union Network of Excellence Epizone EPIZONE (contract no. FOOD-CT-2006-016236) and by the Department for Environment, Food and Rural Affairs.

Dr Rowlands studied for her PhD at the Institute for Animal Health Pirbright Laboratory on the interaction of ASFV with the tick vector *Ornithodoros erraticus* and on the molecular epidemiology of ASFV. She is currently conducting postdoctoral work on the function of ASFV genes involved in inhibiting interferon responses.

References

1. Dixon LK, Escribano JM, Martins C, Rock DL, Salas ML, Wilkinson PJ. Asfarviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger, U, Ball LA, editors. Virus taxonomy, VIIIth Report of the International Committee on Taxonomy of Viruses. London: Elsevier/Academic Press; 2005. p. 135–43.
2. African swine fever, Mauritius. Promed. 2007 Oct 20 [cited 20 Oct 2007]. Available from http://www.afriquenligne.fr/news/daily-news/thousands-of-pigs-killed-in-mauritius-2007102010912
3. Bastos AD, Penrith ML, Cruciere C, Edrich JL, Hutchings G, Roger F, et al. Genotyping field strains of African swine fever virus by partial p72 gene characterisation. Arch Virol. 2003;148:693–706. DOI: 10.1007/s00705-002-0946-8
4. Boshoff CI, Bastos AD, Gerber LJ, Vosloo W. Genetic characterisation of African swine fever viruses from outbreaks in southern Africa (1973–1999). Vet Microbiol. 2007;121:45–55. DOI: 10.1016/j.vetmic.2006.11.007
5. Lubisi BA, Bastos AD, Dwarka RM, Vosloo W. Molecular epidemiology of African swine fever in East Africa. Arch Virol. 2005;150:2439–52. DOI: 10.1007/s00705-005-0602-1
6. African swine fever, Georgia. Promed. 2007 Jun 7 [cited 2007 Jun 7]. Available from http://www.promedmail.org, archive no. 20070607.1845.
7. African swine fever, Georgia Abkhazia Autonomous Republic. Promed. 2007 Aug 21 [cited 2007 Aug 19]. Available from http://www.promedmail.org, archive no. 20070821.2737.
8. African swine fever, Armenia. Promed. 2007 Aug 25 [cited 2007 Aug 24]. Available from http://www.promedmail.org, archive no. 20070825.2793.
9. African swine fever, Nagoro Karabagh. Promed. 2007 Nov 4 [cited 2007 Nov 3]. Available from http://www.promedmail.org, archive no. 20071104.3589.
10. African swine fever, Russia Chechnya. Promed. 2008 Jan 29 [cited 2008 Jan 27]. Available from http://www.promedmail.org, archive no. 20080129.0370.
11. Bastos ADS, Penrith ML, Macome F, Pinto F, Thomson GR. Co-circulation of two genetically distinct viruses in an outbreak of African swine fever in Mozambique: no evidence for individual co-infection. Vet Microbiol. 2004;103:169–82. DOI: 10.1016/j.vetmic.2004.09.003
12. Penrith ML, Pereira CL, Da Silva M, Quembo C, Nhamusso A, Banze J. African swine fever in Mozambique: review, risk factors and considerations for control. Onderstepoort J Vet Res. 2007;74:149–60.
13. Hutchings GH, Ferris NP. Indirect sandwich ELISA for antigen detection of African swine fever virus: comparison of polyclonal and monoclonal antibodies. J Virol Methods. 2006;131:213–7. DOI: 10.1016/j.jviromet.2005.08.009
14. King DP, Reid SM, Hutchings GH, Grierson SS, Wilkinson PJ, Dixon LK, et al. Development of a TaqMan PCR assay with internal amplification control for the detection of African swine fever virus. J Virol Methods. 2003;107:53–61. DOI: 10.1016/S0166-0934-(02)00189-1
15. Nix RJ, Gallardo C, Hutchings G, Blanco E, Dixon LK. Molecular epidemiology of African swine fever virus studied by analysis of four variable genome regions. Arch Virol. 2006;151:2475–94. DOI: 10.1007/s00705-006-0794-z
16. Irusta PM, Borca MV, Kutish GF, Lu Z, Caler E, Carrillo C, et al. Amino acid tandem repeats within a late viral gene define the central variable region of African swine fever virus. Virology. 1996;220:20–7. DOI: 10.1006/viro.1996.0281
17. Saitou N, Nei M. The neighbor-joining method. A new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25.
18. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007;24:1596–9. DOI: 10.1093/molbev/msm092
19. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20. DOI: 10.1007/BF01731581
20. Felsenstein J. Evolutionary trees from DNA sequences. A maximum liklehood approach. J Mol Evol. 1981;17:368–76. DOI: 10.1007/BF01734359
21. Gonzague M, Roger F, Bastos A, Burger C, Randriamparany T, Smondack S, et al. Isolation of a non-haemadsorbing, non-cytopathic strain of African swine fever virus in Madagascar. Epidemiol Infect. 2001;126:453–9. DOI: 10.1017/S0950268801005465
22. Roger F, Ratovonjato J, Vola P, Uilenberg G. *Ornithodoros porcinus* ticks, bushpigs, and African swine fever in Madagascar. Exp Appl Acarol. 2001;25:263–9. DOI: 10.1023/A:1010687502145

Address for correspondence: Linda K. Dixon, Institute for Animal Health, Pirbright Laboratory, Ash Rd, Pirbright, Woking, Surrey GU24 ONF, UK; email: linda.dixon@bbsrc.ac.uk

L'analyse des séquences de 4 gènes du virus PPA nous a permis de caractériser la souche responsable de l'épidémie qui est survenu en 2007 en Géorgie. La comparaison des deux isolats géorgiens avec des isolats en provenance d'autres régions du monde a permis de conclure à une probable introduction unique dans le Caucase, cette conclusion pouvant cependant être tempérée par le faible nombre d'isolats régionaux inclus dans l'étude. Les analyses phylogénétiques du gène codant pour la protéine VP72 ont classé les isolats géorgiens dans le génotype II, avec des isolats en provenance de Madagascar, du Mozambique et de Zambie. Les gènes E183L et CP204L vont plus loin, en rapprochant les isolats géorgiens de 4 souches malgaches isolées entre 1999 et 2003 et de deux souches mozambicaines. Enfin, le gène B602L rapproche lui aussi les souches géorgiennes des mêmes isolats malgaches, et de souches mozambicaines responsables d'épidémie au cours des années 1960 et en 1998. L'hypothèse d'une introduction de la maladie dans le Caucase suivant les routes commerciales maritimes en provenance d'Afrique de l'Est a été envisagée.

La caractérisation des souches responsables de nouveaux foyers épidémiques ou de la récurrence de foyers plus anciens prend ici tout son sens. Nous venons de le voir, la souche virale à l'origine de l'épidémie partie de Géorgie en 2007 et qui a désormais atteint l'ensemble du Caucase, la Fédération de Russie ainsi que, dernièrement, l'Ukraine, se situe dans le même cluster que les isolats malgaches ou mozambicains. De fait, une des explications possibles de l'émergence de la PPA dans cette région, est qu'elle fait suite à des échanges commerciaux par bateau entre la Géorgie et Madagascar ou l'Afrique de l'est. Ce ne sont pas les produits commerciaux eux-mêmes qui sont en cause, mais probablement les déchets de cuisine contenant de la viande de porc contaminée, embarquée au port de départ, distribués ensuite à l'arrivée à des porcs situés à proximité immédiate du port d'arrivée. Compte-tenu de la période estimée d'entrée du virus en Géorgie (fin 2006 - début 2007), l'hypothèse d'un cargo chargé de litchis provenant de Madagascar au cours de l'hiver 2006, a notre préférence. En effet, Madagascar est le troisième exportateur mondial de litchis, mais le premier à destination de l'Europe. Les litchis de Madagascar ont l'énorme avantage d'être récoltés et d'arriver sur le marché international durant les fêtes de fin d'année, au moment où ils sont bien évidement les plus recherchés.

L'utilisation de la séquence nucléotidique du gène codant pour la MCP du virus PPA a permis la classification des isolats viraux en génotypes, ceci permettant d'en connaitre l'origine géographique. Or, l'analyse d'un seul gène ne permet pas une discrimination à l'échelle locale. Une étude plus fine de la phylogénèse du virus, de la même manière que le nombre de souches intégrées aux analyses phylogénétiques permet une meilleure clustérisation des isolats, nécessite l'analyse de plus de gènes afin de mieux caractériser les souches virales. Ainsi, le décryptage de la phylogénèse du virus PPA participe-t-il à une meilleure compréhension de l'épidémiologie de la maladie et donc à son contrôle.

Dans la suite de ce travail, nous avons voulu évaluer si une reconstruction phylogénétique approfondie, utilisant les outils les plus puissants pour analyser les forces

évolutives à l'œuvre dans la diversité virale, c'est-à-dire les méthodes probabilistes de coalescence utilisant le maximum de vraisemblance et l'inférence bayésienne sur plusieurs gènes permettait un décryptage plus fin des liens unissant les virus entre eux. C'est l'objet du troisième volet de ce travail.

Partie 3

De l'étude phylogénétique exhaustive des données de séquences disponibles à l'inférence chronologique de l'origine du virus et de ses clades.

Au-delà de la caractérisation des souches, qui fournit de nombreuses données épidémiologiques sur les souches circulantes de virus PPA, l'analyse des forces qui sous-tendent l'évolution du virus, en inférant son passé, son histoire, peut nous aider à comprendre l'émergence de nouveaux variants. Comprendre l'origine et la propagation de la diversité virale pourrait être d'une grande aide dans l'établissement d'un vaccin efficace contre les différentes souches virales. L'étude de l'évolution du virus PPA, de la façon dont les isolats interagissent entre eux au niveau moléculaire, requiert l'application des méthodes les plus justifiées sur le plan hypothèse d'évolution, telles que le maximum de vraisemblance et l'inférence bayésienne. Pour soutenir nos résultats, trois gènes ont été spécifiquement choisis pour les études phylogénétiques que nous avons menées. Le gène B646L codant pour la protéine de capside VP72, et les gènes E183L et CP204L codant pour les protéines d'enveloppe p54 et p32, qui induisent tous les trois des anticorps contre la réinfection par une souche homologue.

MATÉRIELS ET MÉTHODES

1- Les données

1-1- Les données publiques

Au cours de cette étude nous avons décidé de nous intéresser à l'évolution du virus de la PPA à travers le monde, sans nous restreindre à l'échelle régionale. Pour ce faire, nous avons utilisé le panel le plus exhaustif d'isolats viraux. Ces isolats ont plusieurs provenances. Les sources les plus abondantes de données de séquences sont les banques publiques de données. Ainsi, la majorité des souches de virus que nous avons utilisées provient de la banque de données GenBank (http://www.ncbi.nlm.nih.gov) et de la banque de données disponible sur le site internet du CISA-INIA (http://webainia.inia.es/cisa/asfv/index.asp).

1-2- Les isolats malgaches

Madagascar a été atteint par le virus de la PPA à la fin des années 1990, aussi, peu de données de séquences sont actuellement disponibles à partir des banques publiques. Au cours de cette étude nous avons isolé 21 souches virales malgaches à partir de rates de porcs domestiques prélevées lors d'évènements épidémiques ayant eu cours dans l'île entre 1998 et 2008. Vingt-et-un échantillons ont été sélectionnés selon leur date et lieu de prélèvement de façon à couvrir l'espace-temps. Après isolement viral sur macrophages alvéolaires de porc, nous avons séquencé des gènes ou fragments de gènes et généré ainsi 39 séquences inédites de virus PPA malgaches : 12 séquences du gène B646L, 10 séquences du gène E183L et 17 séquences du gène CP204L.

Pour d'avantage de détails sur les isolats que nous avons utilisés au cours de cette étude, se reporter à l'annexe 1.

1-2-1- Préparation des macrophages alvéolaires

Les macrophages alvéolaires sont extraits à partir de poumons de porc par lavages au PBS 1X. Les alvéoles pulmonaires n'étant pas un milieu stérile, le PBS 1X est supplémenté de pénicilline (20 U/ml), de streptomycine (20 µg/ml) et d'Amphotericin B (0,5 µg/ml) (Gibco) afin de limiter le développement tant des bactéries que des champignons. Après culottage des cellules, les macrophages sont remis en suspension dans du milieu de culture puis les

cellules vivantes sont énumérées dans une cellule de Malassez avec le colorant vital d'exclusion bleu trypan. Les cellules sont ensuite mises en culture à la concentration de 4,5 x 10^5 cellules par cm², soit 8,5 x 10^5 cellules par puits en plaque 24 puits (Falcon). Les macrophages sont cultivés à +37°C en présence de 5% de CO_2, dans du milieu MEM contenant des sels de Earle et de l'HEPES (Eurobio) supplémenté de sérum de fétus de bovin (SFB) ainsi que de pénicilline (20 U/ml) de streptomycine (20 µg/ml) et d'Amphotericin B (0,5 µg/ml) (Gibco).

1-2-2- Isolement viral

Après une nuit d'incubation, le milieu de culture est retiré et les cellules sont lavées dans du PBS 1X. L'isolement viral est effectué au moyen d'homogénats de rates clarifiés dans du milieu MEM. Les macrophages alvéolaires de porc sont inoculés avec 200 µl de surnageant d'homogénat clarifiés puis mis en incubation à 37°C sous atmosphère de 5% de CO_2 pendant une heure, avec agitation de la plaque toutes les 15 minutes. L'inoculum est alors retiré et remplacé par du milieu de culture MEM plus sels de Earle et HEPES frais contenant 5% de SFB et supplémenté de pénicilline (20 U/ml), de streptomycine (20 µg/ml) et d'Amphotericin B (0,5 µg/ml) (Gibco). Après 5 jours d'incubation, les cultures virales sont congelées et décongelées deux fois à -80°C. Le milieu de culture est ensuite transféré, clarifié à 3000 g pendant 10 minutes. Le surnageant de culture est alors conservé à -80°C jusqu'à utilisation.

1-2-3- Purification de l'ADN viral

L'ADN du virus PPA a été purifié à partir de 200 µl de surnageant clarifié de culture au moyen du Kit GFX Genomic Blood DNA Purification Kit (Amersham Biosciences) et selon le protocole prévu par le fabriquant pour l'extraction de l'ADN à partir de sang total. L'ADN purifié a été élué des colonnes de purification dans 100 µl d'eau ultra pure stérile.

1-2-4- Amplification des gènes viraux

La réaction de polymérisation en chaine a été effectuée avec la polymérase FailSafe (Tebu Epicentre). Cette polymérase a la caractéristique d'être très fidèle au brin matrice avec un taux d'erreur d'incorporation des nucléotides lors de la réplication de 3,3 x 10^{-5}. Les gènes cibles ont été amplifiés grâce aux amorces spécifiques suivantes : VP72-d (5'-GGCACAAGTTCGGACATGT-3') et VP72-U (5'-GTACTGTAACGAAGCAGCACAG-3') pour le gène

B646L, E183Lfor_p54 (5'-GGTTGGTTTTCAAATGTTGGCGAAGGTA-3') et E183Lrev_p54 (5'-CCATAAATTCTGTAATTTCATTGCGCCACAAC-3') pour le gène E183L, et p30/32-P1 (5'-TGCCAAGCATACATAAGTTG-3') et p30/32-P2 (5'-ATTTTGCTGTTTATGAATCC-3') pour le gène CP204L. Le milieu réactionnel contenait 100 ng d'ADN total extrait, 0,5 µM de chaque amorce et 2,5 U de polymérase, complété à 25 µL avec de l'eau ultra pure stérile. Le mélange réactionnel a ensuite été additionné de 25 µL de tampon PCR FailSafe C 2X.

Les réactions de polymérisation ont été réalisées comme suit : un cycle de dénaturation de 95°C pendant 5 minutes suivi de 30 cycles comprenant une dénaturation de 30 secondes à 95°C suivie d'une hybridation de 30 secondes à 50°C pour les gènes B646L et CP204L et 55°C pour le gène E183L et enfin, une élongation à 72°C de 30 secondes pour le gène B646L, de 45 secondes pour le gène E183L et de 50 secondes pour le gène CP204L. A la fin des 30 cycles, une élongation finale de 3 minutes à 72°C a alors été réalisée pour terminer la réaction.

La taille des produits d'amplification a été révélée par illumination aux rayons UV après une électrophorèse dans un gel d'agarose à 0,8% en tampon Tris Acétate EDTA (TAE) 1X contenant du bromure d'éthidium (BET), contre un marqueur de poids moléculaire de 100 pb (Biolabs). Les bandes correspondant aux tailles attendues ont ensuite été découpées dans le gel d'agarose puis purifiées aux moyens du kit GFX PCR DNA and Gel Band Purification Kit (Amersham Bioscience) selon le protocole du fournisseur. Les amplicons ont été élués de la colonne de purification dans 25 µl d'eau ultra pure stérile et ont été directement utilisés pour le clonage. Il est à noté que l'illumination de l'ADN sous rayonnement UV a été la plus brève possible afin de limiter la formation de dimères de bases pyrimidines.

1-2-5- Clonage T-A des amplicons PCR

Les extrémités des produits PCR générés par la polymérase du kit PCR FailSafe sont de deux types : l'activité endonucléasique de l'enzyme génère des extrémités franches d'ADN, tandis que son activité 3' exonucléasique accroche en fin d'élongation une adénine non comprise dans la matrice à l'extrémité 3' de l'amplicon. Cette adénine supplémentaire permet de liguer de façon cohésive l'amplicon dans le plasmide pCR2®.1-topo (Invitrogen), plasmide ouvert dont les extrémités 3' terminales sont pourvues d'une thymine non appariée. Outre la thymine cohésive à son extrémité 3', ce plasmide dispose d'une topoisomérase fixée à chaque extrémité qui permettra la ligation de l'amplicon. La ligation a ainsi été réalisée dans un mélange réactionnel contenant 50 ng de plasmide, l'insert (1:3), 1 µl de tampon 6X du kit TOPO TA Cloning® (Invitrogen), complété à 6 µl avec de l'eau ultra pure stérile et incubé pendant 5 minutes à température ambiante. La réaction enzymatique a ensuite été stoppée en plaçant le mélange dans de la glace.

1-2-6- Transformation des bactéries

Suivant la réaction de ligation, des bactéries *Escherichia Coli* Top10 (Invitrogen) chimio-compétentes ont été transformées par choc thermique avec 2 µl de la réaction de ligation. Les bactéries mélangées à l'ADN ont d'abord été incubées dans de la glace pendant 30 minutes. Elles ont alors reçu un choc thermique à 42°C pendant 30 secondes puis ont été replacées dans de la glace pendant 2 minutes. Le mélange a alors été complété avec 250 µl de milieu SOC à température ambiante et placé en incubation à +37°C pendant une heure sous agitation (220 t/m). Les bactéries transformées ont alors été ensemencées sur un milieu gélosé Lysogeny Broth (LB) contenant 30 µM d'ampicilline, 0,2 mM de X-Gal ainsi que 0,5 µM d'IPTG. Les boites de Pétri ainsi ensemencées ont alors été incubées pendant une nuit à +37°C.

1-2-7- Sélection des clones bactériens transformés

Le site d'insertion du plasmide pCR2®.1-topo se situe au milieu du gène Lac-Z codant pour la béta-galactosidase, une enzyme capable de métaboliser le X-Gal. Le produit de métabolisation du X-Gal, couplé à l'IPTG présent dans le milieu de culture entraine la formation d'une coloration bleue des bactéries. Ainsi, si le plasmide s'est refermé en intégrant l'insert, le gène, interrompu, ne pourra coder pour une enzyme fonctionnelle et les colonies bactériennes resteront blanches. En revanche, si le plasmide s'est refermé sur lui-même au cours de la ligation, le gène sera fonctionnel et les colonies bactériennes seront bleues. La première sélection des clones bactériens se fait donc en fonction de la coloration blanc/bleu des colonies. La seconde étape de sélection est faite d'un point de vue moléculaire. En effet, il convient de s'assurer que l'insert intégré dans le plasmide corresponde à l'amplicon PCR produit. Pour le vérifier, des colonies blanches ont été sélectionnées pour effectuer une réaction d'amplification génique de leur insert au moyen des amorces spécifiques décrites en 1-2-4-. Chaque colonie individuelle a été plongée directement dans un milieu réactionnel contenant 2,5 U de polymérase (QIAGEN), 1 µM de chaque amorce, 200 µM de chaque dNTP, 2,5 µl de tampon 10X et de l'eau ultra pure stérile (qsp 25 µl). Le milieu réactionnel a alors été soumis à 5 minutes à 95°C afin de faire éclater les bactéries pour libérer l'ADN, puis l'amplification génique s'est déroulée selon les 35 cycles suivants : 30 secondes de dénaturation à 95°C, 30 secondes d'hybridation à 50°C pour les gènes B646L et CP204L et 55°C pour le gène E183L, et une élongation de 30 secondes à 72°C pour tous les gènes. Une élongation finale de 7 minutes à 72°C a enfin été réalisée pour terminer la réaction.

Les produits d'amplification ont été visualisés après électrophorèse en gel d'agarose à 1% (Qbiogen) en TAE 1X et illumination au rayonnement UV. La taille des amplicons a été

contrôlée en regard d'un marqueur de poids moléculaire 100 pb (Biolabs). Les colonies positives, c'est-à-dire pour lesquelles la réaction d'amplification génique a produit un fragment d'ADN de la taille attendue, ont été sélectionnées afin d'être cultivées et multipliées afin de produire une large quantité de plasmide contenant les gènes ou fragments de gène d'intérêt qui serviront au séquençage. Les plasmides sont de fait plus faciles d'utilisation et permettent souvent un séquençage de meilleure qualité, notamment en ce qui concerne les extrémités de l'insert.

1-2-8- Préparation de l'ADN plasmidique

Chaque colonie bactérienne blanche sélectionnée a été cultivée dans 4 ml de milieu LB contenant 100 µg/ml d'ampicilline. Après une nuit d'incubation à +37°C et 250 t/m les cultures bactériennes ont été centrifugées à 5500 g pendant 15 minutes, et le surnageant de culture a été jeté. L'ADN plasmidique a alors été purifié au moyen du kit Pure Yield Plasmid Miniprep System (Promega), selon les instructions du fabriquant. L'ADN plasmidique purifié a été dosé par spectrophotométrie à 260 nm et son niveau de pureté contrôlé grâce au ratio des DO à 260 et 280 nm.

La présence et la qualité de l'insert présent dans les plasmides ont été contrôlées au moyen d'une digestion enzymatique par EcoRI, dont le plasmide pCR2®.1-topo contient deux sites qui flanquent le site d'insertion des produits PCR et permettent donc son relargage. Le milieu réactionnel contenait 500 ng de plasmide, 10 U d'enzyme EcoRI (Biolabs), 2 µl de tampon 10X et de l'eau ultra pure stérile (qsp 20 µl). Le mélange a été incubé pendant une heure à +37°C, puis les produits de digestion ont été visualisés après électrophorèse en gel d'agarose 1% (Qbiogen) en TAE 1X contre le plasmide non digéré. La taille des fragments d'ADN digérés a été contrôlée après illumination au rayonnement UV au moyen des marqueurs de poids moléculaire 100 pb et 1 kb.

1-2-9- Séquençage des gènes d'intérêt

La méthode qui a été appliquée pour le séquençage des gènes cibles est celle du Big Dye Terminator (ABI PRISM Big Dye Terminator, Perkin Elmer). Le principe de cette méthode est d'utiliser des nucléotides marqués et non marqués au cours de la réaction d'amplification génique. Lorsque les nucléotides marqués sont intégrés au fragment d'ADN néo-synthétisés, leur marquage a un effet terminateur de chaine, la polymérase ne pouvant plus accrocher le nucléotide suivant. Les nucléotides marqués ou non marqués étant recrutés au hasard par la polymérase, les fragments d'ADN générés par l'amplification seront donc de toutes tailles, et leurs extrémités détectables grâce au nucléotide marqué qui les termine. La détection de ce nucléotide final est faite au moyen du séquenceur automatique Perkin Elmer 377, et chaque base a été lue au moins deux fois, en utilisant des amorces

spécifiques du plasmide pCR2®.1-topo qui encadrent l'insert. Les nucléotides ont donc été lus une fois dans le sens 5' → 3' et une fois dans le sens 3' → 5'.

Le milieu réactionnel contenait 600 à 800 ng d'ADN plasmidique, 5 µl de tampon Big Dye Reaction Mix, 1,7 µM d'une seule amorce, 0,75 µl de DMSO, 3 µl de tampon 5X et de l'eau ultra pure stérile (qsp 15 µl). L'amplification génique suivante a alors été réalisée : une dénaturation de 2 minutes à 95°C suivie de 25 cycles composés d'une dénaturation à 95°C pendant 10 secondes, une hybridation de 10 secondes à 50°C et une élongation de 4 minutes à 60°C. Les produits d'amplification ont alors été purifiés sur des colonnes de sépharose (Qbiogene), déshydratés dans un évaporateur et les culots d'ADN repris dans 5 µl d'un mélange bleu Dextran (Perkin Elmer) / formamide (1:5). Avant électrophorèse en gel de polyacrylamide (Cambrex), l'ADN a été dénaturé à 95°C pendant 5 minutes puis placé dans de la glace.

Les séquences ainsi produites ont été contrôlées au moyen du logiciel Vector NTI (Informax Inc.) en utilisant les électrophorégrammes générés par le séquenceur. Après vérification de chaque base, lue au moins deux fois, les fragments complets ont été assemblés. Les nouvelles séquences de gènes du virus PPA ont alors été téléchargées dans une base de données spécifique du virus PPA.

Parmi les 21 isolats malgaches que nous avons séquencés, 19 l'ont été selon cette méthode. Pour les deux derniers, Gara08 et Tsididy08, le séquençage a été externalisé et effectué par la société Beckman Coulter Genomics (Angleterre).

1-3- Création d'une base de données dédiée au virus PPA

Les bases de données publiques de séquences ADN sont rarement dédiées à un seul organisme et, si une telle base de données existe pour le virus Influenza aviaire dans GenBank, aucun outil de ce type n'était jusqu'ici disponible concernant le virus PPA. Nous avons donc créé une base de données permettant d'accueillir de façon exhaustive toutes les séquences disponibles du virus de la PPA. Cette base a été créée au format MySQL et est accessible par le biais d'internet via une adresse URL.

Cette base de données a été abondée de toutes les séquences disponibles dans les banques publiques (GenBank, EMBL, CISA-INIA) ainsi que des séquences que nous avons générées en propre. Douze champs d'informations ont été complétés pour chaque isolat, selon les informations disponibles dans les bases de données et les publications auxquelles elles font références. Ces 12 champs sont : le numéro d'accession, le nom, l'année, le pays, la région de prélèvement et le génotype de l'isolat, l'espèce animale sur laquelle le prélèvement a été effectué, le gène, la taille de la séquence nucléotidique, la séquence ADN, la séquence protéique traduite ainsi que la propriété hémadsorbante du virus.

La base de données a été construite pour sélectionner les isolats à étudier et générer les fichiers de sortie correspondant au format informatique « fasta ». Ces fichiers peuvent alors être directement utilisés dans nombre de logiciels de traitement et d'analyse de séquences.

Au total 361 séquences du gène B646L, 252 séquences du gène E183L et 129 séquences du gène CP204L ont été téléchargées dans la banque de données en ce qui concerne les gènes cibles que nous avons utilisés dans cette étude.

2- Comprendre les relations qui unissent les isolats viraux : analyse approfondie de la phylogénie du virus PPA

2-1- Analyse des données

2-1-1- Alignements

Les alignements multiples des séquences nucléotidiques et protéiques ont été réalisés au moyen du logiciel Mega version 5 (Tamura *et al.* 2011) par la méthode d'alignement progressif ClustalW (Thompson *et al.* 1994). Cette méthode dite heuristique parce qu'elle propose un alignement réalisable mais pas nécessairement optimal, vise à regrouper progressivement les séquences en procédant en trois étapes : l'algorithme calcule tout d'abord l'ensemble des appariements possibles de séquences afin de générer une matrice de distance représentant la divergence nucléotidique ou protéique entre chaque paire de séquences. La matrice de distance sert à construire un arbre guide dont dépendra l'ordre d'incrémentation des nouvelles séquences dans l'alignement, à savoir de la plus similaire à la plus distante. Les deux séquences les plus semblables seront regroupées puis considérées comme une séquence unique qui servira alors de base pour regrouper une à une les autres séquences, ou groupes de séquences identiques.

La matrice servant à la construction de l'arbre guide est une matrice de distances « observées » entre les séquences. Elle a été calculée au moyen d'une méthode d'approximation autorisant l'alignement d'un grand nombre de séquences, car peu gourmande en ressources informatiques et donc en temps (Bashford *et al.* 1987). Cette méthode calcule le nombre de résidus nucléotidiques (ou protéiques) différents entre deux séquences et le divise par le nombre total de sites analysés, conférant ainsi un « score » de distance entre les séquences deux à deux.

A partir de cette matrice de distances, l'algorithme va construire l'arbre phylogénétique guide par la méthode du Neighbor Joining (ou du plus proche voisin) (Saitou & Nei 1987). L'arbre généré a donc des branches dont la longueur est proportionnelle à la

distance observée. La racine de l'arbre est placée de manière à ce que la moyenne des longueurs de branches de chaque côté d'elle soit égale. Chaque séquence de l'arbre peut donc se voir assigner un poids, dépendant de sa distance à la racine, et qui sera ensuite utilisé pour incrémenter les séquences dans l'alignement.

Si cette méthode d'alignement progressif produit des alignements proches de l'alignement optimal, elle ne respecte cependant pas l'ordre des nucléotides dans les codons, ce qui peut amener à rompre le cadre de lecture, par exemple en créant des insertions au sein de codons. Il a donc fallu vérifier manuellement chacun des alignements afin de contrôler le bon respect du cadre de lecture et préserver la réalité biologique de l'alignement. Cette vérification a été réalisée à l'aide du logiciel Mega version 5. Chaque alignement a été traduit en acides aminés et les séquences ont été replacées manuellement dans le cadre de lecture des gènes. De plus, les trous ont été retirés des alignements afin de ne pas induire de divergence artificielle entre les séquences. Au cours de cette vérification, il a été constaté que les séquences disponibles dans les banques de données publiques contiennent certaines erreurs de séquençage non répertoriées, telles que des codons stop. La banque de données ainsi que les alignements ont donc été expurgés de toutes les séquences erronées. Les analyses suivantes ont donc été réalisées à partir de 356 isolats pour le gène B646L, 251 pour le gène E183L et 123 pour le gène CP204L.

2-1-2- Analyse des alignements

2-1-2-1- Saturation des substitutions

Après vérification des alignements, la pertinence de l'information génétique contenue dans le jeu de données a été contrôlée à l'aide du logiciel DAMBE version 5.2.0.14 (Xia & Xie 2001). La valeur informative du jeu de données dépend du phénomène de saturation des substitutions. Ce phénomène survient lorsqu'il devient impossible de discerner si les similitudes de nucléotides observées à un site donné entre deux séquences sont de réelles homologies ou seulement dues au hasard. Dans ce cas, les différences entre les séquences d'ADN ne permettent plus d'analyser les processus évolutifs associés ni d'estimer le temps de divergence entre deux séquences. La méthode utilisée par le logiciel DAMBE est basée sur une analyse de la nature des substitutions observées dans les alignements. Dans la réalité biologique, les transitions représentent la majorité des substitutions nucléotidiques. Ainsi, lorsque la distance génétique entre deux séquences augmente, le nombre de transitions et de transversions augmente proportionnellement, le nombre de transitions étant toujours supérieur. Cependant, dans le cas de séquences de plus en plus éloignées, la saturation des substitutions peut être atteinte et les transversions devenir plus nombreuses que les transitions. Ce phénomène est de fait biologiquement étayé puisque les transversions sont au nombre de huit alors que les transitions sont seulement au nombre de quatre. Dans ce

cas, deux séquences identiques à un site donné ou différant par une simple transition peuvent cependant être séparées de plusieurs évènement de substitutions, le signal phylogénétique des séquences est alors perdu.

2-1-2-2- Détection des recombinaisons

La saturation des substitutions n'est pas le seul biais qui puisse altérer une analyse phylogénétique. Il a été montré que les virus recombinent leurs génomes (Santti *et al.* 1999). Or, la recombinaison est une force évolutive majeure dans l'émergence de variants alléliques. Lors de recombinaisons, des gènes ou des fragments de gènes d'organismes individuels sont échangés. Les génomes qui en résultent sont alors les produits de plusieurs histoires évolutives, ce qui va impacter lourdement les analyses phylogénétiques. En effet, un arbre phylogénétique établissant les relations entre taxons sous le modèle strict de bifurcation ne peut représenter fidèlement l'histoire évolutive d'un génome dont les parties appartiennent à différentes histoires évolutives. L'impact sur la phylogénie sera alors de plusieurs ordres. Dans l'arbre, le ratio entre les longueurs de branches internes et externes va tendre à se réduire et le TMRCA va lui aussi tendre à diminuer (Schierup & Hein 2000a). Enfin, le ratio substitutions non synonymes/substitutions synonymes (dN/dS) sera surestimé, conduisant à détecter de faux sites nucléotidiques soumis à une pression de sélection positive (Shriner *et al.* 2003).

Il convient donc de chercher à détecter tout signal de recombinaison entre des séquences au sein d'un alignement afin d'éliminer les recombinants qui fausseraient l'analyse phylogénétique. Cette détection a été réalisée avec le logiciel RDP3 version 3 (Heath *et al.* 2006). Les tests de détection ont été effectués avec les méthodes RDP (Martin & Rybicki 2000), GENECONV (Padidam *et al.* 1999), MAXCHI (Smith 1992) et SISCAN (Gibbs *et al.* 2000) appliquées à des séquences d'ADN linéaires. Ces méthodes, appelées méthodes de distribution des substitutions, reposent sur une mesure de l'apparentement entre des séquences. Elles identifient les écarts de patrons entre des sites nucléotidiques partagés par des groupes de séquences au sein d'un alignement, en utilisant une base statistique qui permet d'exprimer les différences d'apparentement entre ces séquences au moyen de partitions différentes de l'alignement. Cela revient à compter les sites nucléotidiques communs et/ou différents entre des paires, des triplets ou des quadruplets de séquences. Les sites de début et de fin des recombinaisons sont ainsi détectés et une *p-valeur* de la distance entre les séquences est calculée.

Les séquences de 17 isolats viraux présentant des recombinaisons ont donc été retirées des alignements du gène E183L. Les analyses suivantes ont donc été effectuées sur un total de : 351 séquences du gène B646L, 123 séquences du gène CP204L et 234 séquences du gène E183L.

2-1-2-3- Composition des alignements

Après avoir expurgé les alignements des séquences erronées et des recombinants, la composition nucléotidique des alignements a été déterminée à l'aide du logiciel Dnasp version 5 (Librado & Rozas 2009). Ce logiciel permet d'analyser le polymorphisme des séquences d'un alignement. Le nombre de sites ségrégatifs (sites pour lesquels au moins une substitution est observée au sein de l'alignement) a été calculé, ainsi que le nombre de transversions et de transitions. Parmi ces mutations, le nombre de substitutions synonymes et non synonymes a également été déterminé. La diversité entre deux séquences (π) a ainsi pu être établie. Pi, correspond à la moyenne des substitutions nucléotidiques entre deux séquences pour chaque site étudié (Lynch & Crease 1990). Le nombre moyen de nucléotides différents entre deux séquences (k) a également été calculé (Tajima 1993).

Les taux de substitutions observés (μ) par site et par an maximum et minimum ont été manuellement déterminés comme suit : soit un site nucléotidique n donné dans un alignement contenant S séquences de N sites. Sur ce site on observe que s séquences montrent la même substitution par rapport à la séquence consensus. Le nombre de substitutions peut alors être égal à 1 si l'on considère que les s séquences ont un ancêtre commun qui possédait cette mutation, ou égal à s si l'on considère que chaque séquence a muté au niveau de ce site n ; 1 est donc le nombre minimum de substitution et s le nombre maximum. En répétant ce calcul pour les N sites, on obtient :

$$\mu min = nombre\ de\ substitutions\ diff\acute{e}rentes\ par\ site\ /N/\ Temps\ (ann\acute{e}es)$$

et

$$\mu max = \frac{nombre\ de\ s\acute{e}quences\ diff\acute{e}rentes\ du\ consensus\ par\ site}{/N/\ Temps\ (ann\acute{e}es)}$$

2-1-2-4- Analyse de la pression de sélection (d_N/d_S)

Pour comprendre comment évoluent les séquences codantes d'un jeu de données, il est nécessaire d'évaluer la forme de sélection à laquelle est soumis le gène qui les porte. Pour ce faire, l'analyse des taux de substitutions synonymes et non synonymes, ainsi que de leur ratio s'avère d'une grande utilité. On appelle d_N la distance non synonyme et d_S la distance synonyme entre deux séquences. d_N est définie comme le nombre de substitutions non synonymes D_n s'étant produites à un site non synonyme N, et d_S comme le nombre de substitutions synonymes D_s s'étant produites à un site synonyme S. Dans le cas où un taux de substitution neutre, c'est-à-dire égal en tout point, s'appliquerait pour chaque site des séquences étudiées, et que le nombre de substitutions serait faible, le ratio N/S donnerait une bonne approximation du ratio d_N/d_S, également noté ω. Cependant, cette méthode ne

permet pas d'analyser des processus évolutifs plus complexes, tel que les substitutions multiples au sein d'un même codon, car le chemin évolutif pour passer d'un codon à un autre est lui-même multiple. C'est pourquoi il convient d'appliquer les corrections apportées par les modèles évolutifs utilisés pour la reconstruction d'arbres phylogénétiques (voir ci-dessous). Lorsque le gène qui porte les séquences étudiées, ou plus précisément la protéine encodée par le gène, est soumis à une pression de sélection positive, ou diversifiante, les substitutions non synonymes vont tendre à s'accumuler car elles entraînent un bénéfice pour l'organisme ; on aura donc $\omega > 1$. Si le gène est soumis à une pression de sélection négative, ou purifiante, les variants portant des mutations non bénéfiques, voire délétères seront progressivement éliminés de la population et on aura $\omega < 1$. Enfin, si aucune sélection ne s'applique, ou sélection neutre, on aura $\omega \approx 1$. Dans la réalité biologique, ω ne dépasse pas 1 (sauf localement), car cela signifierait que la pression de sélection positive s'applique de manière constante et infinie sur la séquence protéique.

L'analyse du ratio d_N/d_S a été effectuée pour les gènes B646L, E183L et CP204L à l'aide de l'algorithme codeml (Yang 1998) intégré au logiciel PAML (Phylogenetic Analysis by Maximum Likelihood) version 4 (Yang 2007). Dans le même temps, chaque codon soumis à une pression de sélection positive a été détecté et répertorié.

Le logiciel PAML 4 nécessite l'implémentation de deux fichiers d'entrée : un arbre phylogénétique, et l'alignement dont il est issu. Si l'alignement peut être facilement contrôlé, l'arbre phylogénétique qui servira de support à l'analyse doit être rigoureusement choisi, pour refléter au mieux l'histoire évolutive du gène. Ainsi, l'arbre directionnel choisi pour chaque gène étudié a été retenu selon la méthode décrite en 2-2.

Le fichier de contrôle de l'analyse, identique pour les trois gènes étudiés se trouve en annexe 2. Les paramètres importants fixés pour ces analyses ont été les suivants :

Runmode = 0, ce qui signifie une évaluation de la topologie de l'arbre phylogénétique chargé dans le logiciel.

Seqtype = 1, qui stipule que les alignements servant à l'analyse ont été réalisés avec des séquences codantes, et dans le cadre de lecture des gènes.

CodonFreq = 2 stipule que la fréquence des codons à l'équilibre (soit la proportion dans les séquences des quatre nucléotides pour un temps d'évolution qui serait infini) est calculée à partir des moyennes des fréquences observées des quatre nucléotides sur les trois positions des codons.

Model = 0. Ce paramètre implique que le même taux ω est appliqué sur l'ensemble des branches de l'arbre.

NSsites = 8. Ce modèle (Yang 2000) prend en compte 11 classes pour les sites nucléotidiques : 10 classes pour la distribution β ainsi qu'une classe supplémentaire pour les

codons ayant une valeur de $\omega \geq 1$, c'est-à-dire soumis à une pression de sélection positive. Ce modèle a été utilisé en parallèle avec la variable ncatG (nombre de catégories de la distribution de ω) fixée à 8.

GetSE a été fixé à 1 afin de permettre une estimation des erreurs standards associées aux paramètres.

Enfin, RateAncestor = 1 qui va permettre à l'algorithme de procéder à une reconstruction de la séquence ancestrale des alignements étudiés par une approche bayésienne empirique, qui utilise les longueurs de branches de l'arbre ainsi que le taux de substitution relatif de chaque nucléotide (Koshi & Goldstein 1996).

L'utilisation de toutes les séquences d'isolats disponibles n'a pas permis de mener à bien ces analyses. En effet, les erreurs standards convergeaient vers zéro avant la fin de l'analyse, indiquant une divergence de l'algorithme ou des longueurs de branches égales à zéro. Des jeux de données réduits, représentant les séquences uniques dans l'alignement ont alors été utilisés. Soit 67 séquences pour le gène B646L, 70 séquences pour le gène E183L et 65 séquences pour le gène CP204L.

2-2- Reconstructions phylogénétiques

2-2-1- Choix du modèle évolutif

Les méthodes de maximum de vraisemblance (Felsenstein 1981) ont pour but de déterminer le modèle évolutif qui correspond le mieux à l'évolution qui s'exprime au sein d'un jeu de données de séquences. Cette détermination passe par l'application de modèles évolutifs probabilistes. Les analyses en maximum de vraisemblance suivantes ont été réalisées à l'aide du logiciel TREEFINDER version mars 2011 (Jobb *et al.* 2004).

En premier lieu, le modèle évolutif le plus approprié aux jeux de données dont nous disposions a été déterminé en utilisant l'option « propose model » du logiciel TREEFINDER. Afin de consolider les propositions faites pour chaque alignement, trois critères d'information ont été employés : Akaike Information Criterion (AIC) (Akaike 1974), qui est une estimation du taux d'information perdue lors de l'utilisation d'un modèle pour représenter un processus stochastique, ce qu'est l'évolution. Les modèles donnant un AIC faible sont les plus appropriés au jeu de données analysé. L'AIC d'un modèle M est égal à :

$$AIC = -2 \ln vraisemblance + 2k$$

avec k le nombre de paramètres estimés.

Le second critère est un dérivé du précédant, appelé AICc (Sugiura 1978) (pour AIC corrigé). Le calcul de l'AICc se fait selon la formule suivante :

$$AICc = -2 \ln vraisemblance + \frac{2km}{m - k - 1}$$

avec k le nombre de paramètres estimés et m le nombre de sites étudiés.

Il est à noter que l'AIC et l'AICc tendent à converger lorsque m grandit.

Enfin, le Bayesian Information Criterion (BIC) (Schwarz 1978) a été utilisé. Ce test pénalise plus les paramètres libres des modèles que ne le font les AIC. Il se calcule comme suit :

$$BIC = -2 \ln vraisemblance + k \ln(m)$$

Ces trois tests ont été réalisés pour une distribution gamma (Γ) des nucléotides partitionnée en 5.

2-2-2- Construction des arbres phylogénétiques

2-2-2-1- Maximum de vraisemblance

Pour chacun des trois gènes analysés, à savoir B646L, CP204L et E183L, de 1 à 3 arbres phylogénétiques ont été construits en utilisant le logiciel TREEFINDER, selon les modèles déterminés en 2-2.1-. Pour des raisons de comparaison, le model réversible généralisé (GTR pour *General Time Reversible*) (Lanave *et al.* 1984 ; Rodriguez *et al.* 1990) a été systématiquement utilisé. Ce modèle est le plus complexe de tous les modèles de reconstructions phylogénétique avec 9 paramètres (les 3 paramètres de fréquences à l'équilibre et les 6 paramètres d'échangeabilité), les autres modèles étant des cas particuliers de ce modèle. Pour renforcer le degré de confiance de chaque arbre construit, deux analyses de ré-échantillonnage utilisant 1000 répétitions ont été effectuées : le test Local – Rearrangements / Expected-Likelihood Weight (LR-ELW) (Strimmer & Rambaut 2002) et le test de bootstrap non paramétrique (Felsenstein 1985). Lors de l'analyse ELW, toutes les topologies possibles autour de chaque branche interne d'un arbre non enraciné sont générées et la longueur des nouvelles branches ainsi créées est mesurée sans que soient modifiés les autres paramètres du modèle. La valeur de vraisemblance de toutes ces topologies réunies est alors calculée et affectée à la branche originelle, en pourcentage.

Au cours de l'analyse avec bootstrap ce ne sont pas les longueurs de branches qui sont testées individuellement, mais un ré-échantillonnage complet des alignements est effectué. La méthode consiste à créer artificiellement de nouveaux alignements, ici 1000, de la même taille que l'alignement originel, en effectuant un tirage au sort avec remise des colonnes de

l'alignement originel. Une reconstruction phylogénétique est alors faite à partir de chaque alignement chimérique généré, et la valeur associée à chaque nœud dans l'arbre correspond au nombre de fois sur 1000 réplicats où cette même topologie a été trouvée. La valeur du bootstrap est indiquée en pourcentage.

La congruence des topologies d'arbres générées a été testée grâce à l'option « Analysis | Test Hypotheses » du logiciel TREEFINDER et le test ELW de Strimmer et Rambaut (2002) a été appliqué. Ce test compare entre eux des arbres construits selon des modèles phylogénétiques différents appliqués au même jeu de données. Une valeur de vraisemblance est calculée et une *p-valeur* est attribuée à chaque arbre. L'arbre ayant la valeur ELW la plus haute a été retenu.

2-2-2-2- Inférence bayésienne

Contrairement aux méthodes par maximum de vraisemblance qui analysent les probabilités *a posteriori*, l'approche par inférence Bayésienne permet de calculer ou de réviser la probabilité *a priori* d'une hypothèse (Holder & Lewis 2003). La méthode prend ainsi en compte cette probabilité postérieure pour calculer la probabilité *a priori* de l'hypothèse suivante. L'analyse phylogénétique par inférence Bayésienne a été effectuée par des chaines de Markov avec technique de Monte Carlo (*Markov Chain Monte Carlo*, MCMC) du logiciel Mr Bayes version 3.1 (Huelsenbeck & Ronquist 2001 ; Ronquist & Huelsenbeck 2003). Les MCMC sont une méthode statistique d'échantillonnage utilisant des fonctions intégrées. Elles permettent de faire des tirages aléatoires d'échantillons (technique Monte Carlo) à partir des fonctions, chaque tirage étant basé directement sur le résultat du tirage précédent (chaine de Markov). Ainsi, les probabilités générées par une MCMC, si elles sont très différentes au début de l'analyse, puisqu'issues d'un tirage aléatoire, finissent par converger, puisque leur expression est basée sur le résultat du tirage précédent.

Une chaine de Markov, ou processus de Markov, considère chaque site d'une séquence d'ADN comme une variable aléatoire dont les différents états n forment une fonction discrète, c'est-à-dire discontinue. Dans le cas de l'ADN, les différents états d'un même site sont au nombre de 4 : les 4 nucléotides A, T, C et G. Ainsi, un processus de Markov permet de définir la probabilité de remplacement d'un nucléotide par un autre après une période de temps t, tout le long des séquences d'ADN étudiées.

Les modèles utilisés ont été ceux qui se sont révélés être les plus adaptés aux jeux de données que nous avons utilisés, c'est-à-dire ceux pour lesquels la valeur du test ELW a été la plus forte lors du test des hypothèses en maximum de vraisemblance. Le modèle GTR a, là aussi, été réalisé pour comparaison. Les chaines de Markov ont été tournées le nombre de cycles nécessaires pour que la valeur du test de vraisemblance (LRT) associé aux arbres générés soit inférieure à 0,01 quand cela était possible, ou du moins stabilisée à la valeur la

plus proche de 0,01. L'arbre consensus a ensuite été défini par comparaison de tous les arbres générés au cours des MCMC après avoir retiré les premiers 25% des arbres générés par l'analyse (burn-in). En effet, lors de sa phase initiale les LRT produits par les MCMC sont élevés, car très influencés par le point de départ de l'analyse. Des LRT élevés signifient des arbres assez voire très dissemblables et reflétant donc peu la réalité du jeu de données. Inclure ces arbres dans le processus de génération de l'arbre consensus pourrait donc en altérer la topologie.

2-2-2-3- Enracinement des arbres

Les méthodes phylogénétiques produisent en général des arbres non enracinés car, si elles sont capables de déterminer les relations entre les séquences des isolats, elles n'ont pas la capacité d'orienter le processus d'évolution dans le temps. Il existe plusieurs méthodes pour enraciner un arbre. La plus simple d'entre elles consiste à placer la racine à mi-chemin entre les taxons les plus dissemblables de l'arbre, c'est la méthode de racinement au barycentre (Faris 1972). Cependant, cette méthode n'est pas complètement satisfaisante, puisqu'elle présuppose que l'horloge moléculaire s'applique aux séquences étudiées, à savoir que tous les sites nucléotidiques ont évolués à la même vitesse au cours du temps (Zuckerkandl & Pauling 1965). La méthode la plus fiable pour enraciner un arbre consiste à inclure dans le jeu de données une ou plusieurs séquences homologues mais distinctes des séquences à analyser, constituant un groupe externe. Outre que cela implique de connaitre les limites du phylum que l'on étudie, il convient de ne pas utiliser des séquences provenant d'organismes trop distants du groupe étudié. Cela aurait pour effet de faire intervenir le phénomène de saturation des substitutions, et les sites perdraient leur valeur informative d'un point de vue phylogénétique.

Le virus de la PPA est le seul membre de la famille des *Asfarviridae*. Les virus sélectionnés pour représenter le groupe externe ont donc été choisis parmi les familles de virus qui ont été trouvées les plus proches des *Asfarviridae* (Iyer *et al.* 2001) : les *Iridoviridae* (famille dans laquelle avait été initialement placé le virus PPA), les *Poxviridae*, et les *Phycodnaviridae*. Au total, quatre virus ont été sélectionnés : le Lymphocystis Disease virus 1 (LDV1) et le Invertebrate Iridescent virus 6 (CIV6) pour la famille des *Iridoviridae*, le virus de la Vaccine (VV) pour les *Poxviridae* et le virus Ectocarpus siliculosus (EsV) pour les *Phycodnaviridae*. Iyer *et al*. avaient basé leur étude phylogénétique de virus à ADN double brins sur l'utilisation de la protéine majeure de la capside virale. Dans le cas du virus PPA, il s'agit du gène B646L. Un alignement multiple des séquences protéiques a donc été produit à partir des séquences protéiques complètes de 31 isolats de virus PPA et 4 de leurs homologues. Les séquences étant relativement divergentes, l'algorithme qui dirige la méthode ClustalW n'est pas parvenu à produire un alignement correct des séquences. L'alignement a donc été réalisé manuellement à l'aide du logiciel Mega version 5. L'arbre

phylogénétique a ensuite été construit selon le processus décrit ci-dessus : le choix du modèle, la construction et la comparaison des hypothèses.

Pour orienter un arbre phylogénétique dans le temps, lorsque l'alignement ne comporte pas de groupe externe, il convient de placer manuellement la racine sur la branche de l'arbre menant au taxon ou groupe de taxa le plus ancien, c'est-à-dire l'ancêtre commun des autres taxa. Or, pour un arbre contenant n taxa, il existe $2n-3$ branches, soit autant de positionnements possible pour la racine. L'arbre qui a été construit en incluant des séquences externes fournira cet emplacement. En effet, le nœud de l'arbre qui lie le groupe externe aux séquences d'intérêt représente l'ancêtre commun de l'ensemble des isolats, puisqu'à partir de ce point les familles virales ont divergé. Il figure donc également l'ancêtre des isolats de virus PPA. La racine devra donc être placée sur la branche qui, partant de ce nœud ancestral joint le groupe d'intérêt.

2-3- Classification des isolats de virus PPA

2-3-1- Approche par l'utilisation des distances entre isolats

La cladistique, ou systématique phylogénétique, est basée sur la recherche des homologies existant entre des organismes afin de déterminer lesquels ont en commun un ancêtre, ancêtre commun impliquant que le groupe de descendants est monophylétique. La construction d'un arbre phylogénétique est basée sur l'analyse des distances existant entre des séquences d'ADN ou de protéines, en traduisant ces distances en longueurs de branches.

Afin de procéder à la classification des isolats dont nous disposions, nous avons analysé la matrice de distance générée lors de la construction de l'arbre ayant obtenu le score le plus élevé au test ELW décrit en 2-2-. Le gène choisi pour cette classification a été le gène B646L car c'est celui qui a été à la base du génotypage actuel des virus PPA (Bastos *et al.* 2003). De plus, la protéine majeure de capside a été trouvée appropriée pour étudier l'évolution et les relations existant entre des virus de la plus proche famille des *Asfarviridae*, les *Iridoviridae* (Tidona *et al.* 1998).

A partir de cette matrice, les fréquences des distances ont été calculées. Ces fréquences caractérisent les relations au niveau intergroupes et au niveau intra-groupe. En effet, les distances entre les clades doivent être plus élevées que les distances au sein d'un même clade de séquences, permettant ainsi de déterminer les frontières de chaque groupe d'isolats viraux. Les moyennes des distances au niveau intergroupes et intra-groupes ainsi que les maximums et minimums ont également été calculés.

Le regroupement initial des isolats a été effectué en se basant sur la classification en génotypes établie pour le gène B646L depuis 2003 par Bastos *et al*., (génotypes I à X), puis complétée par Lubisi *et al*. en 2005 (génotypes XI à XVI) et Boshoff *et al*. en 2006 (génotypes XVII à XXII). Pour chaque groupe prédéfini, le maximum de distance interne a été déterminé et comparé avec le minimum de distance entre deux groupes. Si le maximum interne s'est révélé être plus élevé que le minimum au niveau intergroupes, les isolats ou groupes d'isolats incriminés ont été replacés au sein de nouveaux sous-groupes, ou sous-génotypes. Les calculs des moyennes, maximums intra-groupes et minimums intergroupes ont alors été calculés pour cette nouvelle configuration. De proche en proche, les isolats ont ainsi été classés.

2-3-2- Approche en réseau

Si les distances calculées entre des séquences nucléotidiques rendent bien compte des homologies et des différences qui existent entre ces séquences, elles n'informent en rien quant à leur généalogie. En effet, un arbre phylogénétique et la matrice afférente ne précisent pas le chemin, en termes de nombre de mutations, nécessaire pour passer d'une séquence à une autre. De même, un arbre phylogénétique ne permet pas qu'un nœud interne soit représenté par un taxon, tous les taxa étant des feuilles de l'arbre, terminant une branche externe. Pour assurer davantage notre classification des isolats de virus PPA, nous avons voulu déterminer si les relations entre les taxa étaient représentées au mieux par un arbre phylogénétique « simple », c'est-à-dire basé sur la bifurcation, et donc présupposant que le processus évolutif sous-jacent est lui-même basé sur la bifurcation, ou par un arbre dont le modèle de construction autorise la représentation de plusieurs chemins conduisant d'une séquence à une autre (polytomie), c'est-à-dire un réseau. Un réseau, parce qu'il n'anticipe pas le modèle évolutif des séquences comme étant un arbre, présente l'avantage de ne pas « forcer » les données à être modélisées comme un arbre. Ceci confère la possibilité d'étudier des processus évolutifs plus complexes, tels que les recombinaisons (mais les isolats recombinants ont été retirés des analyses), les transferts latéraux de gènes et les hybridations (processus évolutifs bactériens plutôt que viraux).

Deux méthodes de construction de réseaux d'haplotypes ont été utilisées.

La première méthode est basée sur une analyse statistique en parcimonie (Templeton *et al.* 1992) utilisée par le logiciel TCS version 1.21 (Clement *et al.* 2000). Cette méthode calcule la probabilité en parcimonie des différences existant entre des séquences, en comparant les séquences deux à deux. Le calcul est poursuivi jusqu'à ce que la probabilité dépasse le seuil de 95%. Les connexions entre les séquences, c'est-à-dire le nombre de substitutions permettant de passer d'une séquence à une autre, correspondent au nombre de mutations associées aux probabilités juste avant qu'elles ne dépassent ce seuil de 95%.

Ainsi, il sera possible de déterminer le nombre de mutations nécessaires pour passer d'un groupe de séquences à un autre, et ainsi établir les limites des dits-groupes.

La seconde méthode est celle des réseaux de partitions ou « *split-networks* ». Elle utilise la décomposition par partitionnement, ou « *split decomposition* » (Bandelt & Dress 1992b ; Bandelt 1992a), du logiciel SPLITSTREE version 4 (Huson & Bryant 2006). La « split decomposition » introduit la notion de partitionnement des données. L'algorithme va alors construire les arbres déterminés par les différents partitionnements et représenter toutes les branches qui lient les taxa entre eux. C'est la raison pour laquelle des branches parallèles apparaissent dans le réseau : elles représentent les distances phénétiques entre des taxons dont le chemin est retrouvé lors de la construction d'arbres différents.

Le réseau de partitionnement a été réalisé en utilisant l'option « *NeighborNet* » du logiciel SPLITSTREE version 4 puis l'option « Bootstrap » a été appliquée avec un nombre de répétitions égal à 100. Le réseau final a été généré à partir de cette analyse de ré-échantillonnage en utilisant l'option « *show bootstrap network* ».

2-3-3- Approche biologique

Le calcul des distances entre des séquences d'ADN est fondé sur la détection des différences, en termes de nucléotides, qui existent entre ces séquences. Or, si ces distances reflètent bien l'éloignement qui sépare des séquences dans un jeu de données, elles ne disent en rien ce qui les réunit. En effet, il a été constaté que la distance maximale entre des séquences d'un même groupe pouvait être très supérieure à la distance minimale entre deux groupes distincts dans l'arbre phylogénétique (cf. résultats). Ainsi, pour comprendre ce qui lie des taxons entre eux, l'alignement du gène B646L a été exploré afin de déterminer la signature moléculaire de chaque groupe distinct de taxa, signature moléculaire à l'origine du regroupement des séquences au sein de clades dans l'arbre, et donc de la possible classification en géno-groupes.

Pour ce faire, l'alignement a tout d'abord été expurgé de toutes les séquences doublons (c'est-à-dire identiques) à l'aide du logiciel DAMBE version 5.2.0.14 (Xia & Xie 2001). Ensuite, les sites conservés ont été retirés manuellement des deux alignements, afin de ne conserver et d'analyser que les sites variables, sièges de la signature moléculaire des isolats, à l'aide du logiciel Mega version 5 (Tamura *et al*. 2011). Une séquence consensus 50 a alors été déterminée au moyen du logiciel seaview version 4-2-7 (Galtier et al. 1996 ; Gouy et al. 2010). Au final, le décryptage des signatures moléculaires des géno-groupes à été réalisé sur un alignement de 67 séquences uniques (sur 351 au total) et 93 sites d'intérêt (sur 399 nucléotides). Les isolats ont alors été regroupés selon les grandes lignées observées après la construction des arbres phylogénétiques, des réseaux d'haplotypes et des réseaux de partitionnements.

Les sites conservés à l'intérieur de ces grandes lignées ont alors été également retirés manuellement des sous-alignements, pour ne laisser apparaitre que les signatures moléculaires spécifiques des grandes lignées elles-mêmes ainsi que des géno-groupes qui les composent.

3- Datation moléculaire

Deux méthodes ont été utilisées pour déterminer le taux d'évolution moléculaire μ ainsi que la date du plus récent ancêtre commun (*Time For The Most Recent Commun Ancestor*, TMRCA) des isolats de virus PPA circulants. Le taux d'évolution moléculaire μ mesure le taux auquel des organismes varient (ou divergent) au fil du temps. Il est exprimé en nombre de substitutions étant survenues dans une séquence par site nucléotidique et par unité de temps (en année). A partir de ce taux d'évolution, l'âge du plus récent ancêtre commun sera inféré. Les deux méthodes utilisées, comme pour la reconstruction phylogénétique, ont été le maximum de vraisemblance et l'inférence bayésienne. Quelles que soient les méthodes utilisées, le résultat du calcul de μ et du TMRCA prend en compte le facteur temps. Pour chacune des séquences de gène intégrées dans l'analyse, la date d'isolement du virus devait donc être connue.

A chaque site d'une séquence d'ADN, le taux de substitution nucléotidique peut ne pas être égal. En effet, une pression de sélection positive telle que celle exercée par le système immunitaire de l'hôte induit logiquement un taux de substitution supérieur pour les codons sur lesquels elle s'applique. Le calcul du taux d'évolution associé à ces codons est donc soumis à un biais de par cette force évolutive, et ne correspond donc pas à l'évolution naturelle des séquences d'ADN de l'organisme étudié. Ce biais aura une incidence sur le calcul du TMRCA, un taux de substitution plus élevé devant entrainer un recul de l'âge du TMRCA.

Pour circonvenir à ce biais d'analyse, les codons soumis à pression de sélection positive détectés en 2-1-2-4- pour chacun des gènes étudiés ont été retirés de leur alignement respectif.

3-1- Datation moléculaire par maximum de vraisemblance

La datation moléculaire par maximum de vraisemblance a été réalisée à l'aide de l'algorithme Baseml qui suit des modèles de substitutions markovien, intégré dans le logiciel PAML version 4 (Yang 2007).

Cet algorithme permet l'utilisation de différents modèles de substitutions nucléotidiques. En cohérence les analyses de reconstruction phylogénétique que nous avons réalisées en 2-2-2-, les modèles que nous avons utilisés pour la datation moléculaire ont été ceux qui ont été trouvés comme les plus appropriés à nos jeux de données de séquences. Ainsi, le modèle HKY + $\Gamma 5$ (Hasegawa *et al.* 1985) a été utilisé pour déterminer μ et le TMRCA des trois gènes étudiés B646L, E183L et CP204L.

Comme pour la détermination du taux de sélection positive, le logiciel PAML version 4 requiert le chargement d'un arbre phylogénétique et de l'alignement ayant servi de base à sa construction. Les arbres ayant été construits selon les modèles cités ci-dessus ont donc été implémentés dans le logiciel.

Pour chaque gène, l'hypothèse de l'horloge moléculaire stricte, ou hypothèse nulle, (Zuckerkandl & Pauling 1965), à savoir un taux d'évolution moléculaire identique pour chaque site nucléotidique tout le long des séquences étudiées a été testée contre une analyse sans qu'aucune horloge moléculaire ne soit appliquée. Deux arbres phylogénétiques ont donc été construits : l'un en utilisant l'option clock=1 (hypothèse nulle) qui a ensuite été comparé avec un arbre construit avec l'option clock=0 (non contraint par une horloge moléculaire, et donc pour lequel μ ne sera pas déterminé). Pour chaque arbre, une valeur de vraisemblance (LR) a été calculée. La valeur de cette probabilité est proportionnelle à celle que l'on pourrait observer pour un jeu de séquences et un modèle probabiliste donné. Cette valeur peut être considérée comme l'expression de l'adéquation entre le modèle suivi et le jeu de données. Ce sont ces valeurs qui sont utilisées pour comparer les arbres phylogénétiques entre eux, au moyen d'un ratio test de vraisemblance (LRT), qui permet de choisir entre les deux hypothèses testées, à savoir l'hypothèse nulle H_0 contre une hypothèse alternative H_1. Si l'hypothèse nulle est vraie, la valeur obtenue par la comparaison des deux hypothèses doit suivre une distribution χ^2 avec un degré de liberté égal au nombre de paramètres indépendants entre les deux modèles, déterminé par le nombre N de séquences dans l'alignement. L'hypothèse H_0 demande d'implémenter un arbre phylogénétique non enraciné, c'est-à-dire pour lequel $2N-3$ branches internes sont estimées alors que l'hypothèse alternative H_1 requiert un arbre raciné, et donc pour lequel $N-1$ branches internes sont estimées. La détermination du LRT est faite selon la formule suivante :

$$LRT = 2(LR\ H_1 - (LR\ H_0))$$

Le nombre de degrés de liberté étant égal à : $2N-3-(N-1)=N-2$.

Le LRT sera alors comparé avec la valeur de χ^2 calculée pour le même nombre de degrés de liberté, ou χ^2critique (χ_c^2). Si le LRT dépasse la valeur du χ_c^2, alors la perte de vraisemblance de l'arbre phylogénétique construit sous l'hypothèse H_0 est significative, et l'hypothèse nulle sera rejetée au seuil de 95%, soit avec un risque de 5%.

Outre les choix des options clock=0 et clock=1, l'option getSe=1 a été choisie afin que les calculs d'incertitudes (*standard errors*) soient effectués lors de l'analyse.

L'horloge moléculaire stricte, c'est-à-dire un taux d'évolution constant au travers de toutes les branches d'un arbre phylogénétique, est biologiquement rarement vérifiée. En effet, les forces évolutives, comme par exemple la pression du système immunitaire de l'hôte, ne s'appliquent pas de la même façon sur l'ensemble d'une séquence nucléotidique. L'option clock=2 a donc été utilisée. Cette option permet en effet de faire varier μ, selon les branches, à l'intérieur de l'arbre. Pour ce faire, les branches supposées diverger à des taux d'évolution différents (possédant donc une horloge « locale ») doivent être signalées et numérotées dans la syntaxe de l'arbre implémenté dans le logiciel PAML version 4 et destiné à servir de guide pour l'analyse. La syntaxe d'un arbre phylogénétique s'écrit comme suit : une branche est symbolisée par une paire de parenthèses ; à l'intérieur de ces parenthèses, les noms des isolats dont les séquences sont étudiées et qui appartiennent à cette branche sont cités, et la longueur de la branches dont ils sont la feuille est indiquée. La longueur des branches partant et menant aux nœuds internes de l'arbre est également indiquée. A titre d'exemple, la syntaxe d'un arbre à $N = 4$ feuilles est la suivante :

$$(N_1 : x_1, ((N_2 : x_2, N_3 : x_3) : x_{2-3}, N_4 : x_4) ;$$

avec x_n la longueur de branche associée à une feuille ou à un groupe de feuilles. La signalisation de la branche contenant, par exemple, les feuilles N_2 et N_3 sera donc effectuée en plaçant le symbole # après la parenthèse fermant la branche qui les contient : $(N_1 : x_1, ((N_2 : x_2, N_3 : x_3)\#1 : x_{2-3}, N_4 : x_4))$;

Les branches autorisées à évoluer avec un taux d'évolution différent ont été déterminées en prenant pour base les branches de l'arbre issu du modèle le plus en adéquation avec le jeu de données correspondant. Des analyses successives ont alors été réalisées, en faisant varier le nombre de branches pouvant avoir un taux d'évolution propre. Dans ces cas, l'analyse par maximum de vraisemblance implémentée dans PAML version 4 ne permet pas d'obtenir un taux d'évolution moléculaire global de l'arbre, mais le taux d'évolution de chaque branche marquée dans l'arbre.

Les fichiers de contrôle ayant servi de base à ces analyses et associés à chaque gène se trouvent en annexe 3.

3-2- Datation moléculaire par inférence bayésienne

La détermination du taux d'évolution ainsi que du TMRCA des gènes B646L, E183L et CP204L du virus PPA par inférence bayésienne a été réalisée à l'aide du logiciel BEAST version

1.6.2 (Bayesian Evolutionary Analysis by Sampling Trees) (Drummond & Rambaut 2007). Les chaines de Markov bayésiennes avec technique de Monte Carlo ont été réalisées selon un processus semblable à celui mis en œuvre avec le logiciel Mr Bayes (Huelsenbeck & Ronquist 2001 ; Ronquist & Huelsenbeck 2003), décrit en 2-2-2-2-. La méthode d'inférence bayésienne utilisée par le logiciel BEAST offre plusieurs avantages en comparaison avec les analyses en maximum de vraisemblance. Elle permet d'intégrer des modèles évolutifs très complexes et son processus est plus rapide. De plus, à la méthode développée dans le logiciel Mr Bayes, et qui met en exergue l'inférence phylogénétique, la méthode implémentée dans le logiciel BEAST additionne des méthodes de coalescence développées pour analyser la génétique des populations : les méthodes LAMARC (Kuhner 2006) et BATWING (Ian J. Wilson 2003).

La phylogénie moléculaire et la coalescence n'ont pas tout à fait le même but. Si la première recherche l'arbre « vrai » décrivant les relations entre des loci géniques homologues, la seconde recherche moins la vérité de l'arbre que la compréhension de la généalogie retraçant l'histoire des séquences étudiées. Pour ce faire, elle vise à décrypter les forces évolutives sous-jacentes à la généalogie de la population analysée. Ces forces comprennent, par exemple, les recombinaisons, le taux de croissance des populations, leur processus de sélection ainsi que leur divergence. En conséquence, la théorie de la coalescence (Kingman 1982) est un ensemble structuré de théories mathématiques visant à déterminer la date à laquelle des séquences ont divergé, c'est-à-dire le temps auquel existait leur ancêtre commun.

L'algorithme qui dirige les analyses effectuée par le logiciel BEAST est une chaine de Markov avec technique de Monte Carlo couplée à l'algorithme de Metropolis – Hasting (Metropolis Coupling Markov Chain Monte Carlo, MCMCMC) (Hastings 1970 ; Metropolis 1953). Cette méthode permet de pallier une des faiblesses d'une MCMC : l'existence de maxima locaux, qui peuvent fausser les approximations d'une chaine markovienne. En effet, ces maxima locaux peuvent stopper une MCMC alors que l'exploration des paramètres n'a pas été totale. La première façon de déjouer ce piège est de faire tourner la MCMC sur un nombre suffisant de générations pour s'assurer de la fiabilité des probabilités postérieures. Cette méthode permet de faire tourner simultanément plusieurs MCMC. Tandis que la première chaine, dite chaine « froide » fournit une estimation des probabilités postérieures en convergeant vers une distribution stationnaire, les autres chaines, dites chaines « chaudes » explorent d'autres versions de la même distribution. A intervalle régulier, l'état des chaines est permuté selon le processus de Metropolis, le nouvel état est alors comparé au précédent, puis accepté ou rejeté selon son adéquation au jeu de données étudié. Les chaines chaudes et froides tournent en parallèle, mais seul l'échantillonnage produit par la chaine froide servira de base à l'inférence.

Ce couplage des MCMC permet d'explorer des processus évolutifs beaucoup plus complexes, comme la modélisation explicite du taux d'évolution moléculaire associé à

chaque branche d'un arbre phylogénétique. Cela permet au logiciel BEAST de produire des arbres racinés possédant une échelle de temps et donc d'étudier des séquences d'un point de vue généalogique. Outre l'hypothèse la plus simple de taux d'évolution fixe tout le long de la séquence (horloge moléculaire stricte), l'algorithme implémenté dans BEAST permet l'utilisation d'horloges moléculaires « relâchées », c'est-à-dire l'autorisation pour le taux d'évolution de varier le long de l'arbre (Kishino *et al.* 2001 ; Thorne & Kishino 2002 ; Yoder & Yang 2000). Contrairement au logiciel PAML version 4 et son option « horloge locale », qui demandait à l'utilisateur de choisir lui-même les branches pouvant avoir un taux d'évolution différent, ici, l'algorithme teste les hypothèses pour toutes les branches, en calculant les probabilités postérieures de l'hypothèse testée qu'il acceptera ou refusera, selon le résultat.

La suite BEAST est un ensemble de logiciels dont l'utilisation progressive permet d'analyser la généalogie d'un jeu de données de séquences d'ADN. Le logiciel BEAUTI permet de spécifier un grand nombre de modèles évolutifs que l'on souhaite appliquer à un jeu de séquences donné. Après que l'analyse ait été effectuée par le logiciel BEAST, les résultats sont analysés à l'aide du logiciel TRACER. L'arbre phylogénétique consensus gradué dans le temps est enfin généré au moyen du logiciel TREEANNOTATOR, qui produit un arbre regroupant toutes les informations contenues dans l'ensemble des arbres générés au cours de l'analyse.

Les modèles évolutifs que nous avons choisis d'utiliser pour explorer nos séquences ont été basés sur ceux qui ont été trouvés les plus en adéquation avec nos jeux de données. Il s'agissait des modèles HKY + $\Gamma 5$ (Hasegawa *et al.* 1985) pour les trois gènes B646L, E183L et CP204L.

L'inférence bayésienne autorise à postuler *a priori* des hypothèses (« *priors* ») autres que le modèle évolutif, et donc de fixer avant l'analyse certains paramètres, hypothèses dont la plausibilité sera vérifiée au moyen de probabilités postérieures.

Ces hypothèses *a priori* ont pour objectif de calibrer les informations afin de pouvoir discerner le taux d'évolution μ du temps dont il dépend. Le processus évolutif qui conduit la diversification des séquences d'ADN du virus PPA a conduit à rejeter l'hypothèse nulle de l'horloge moléculaire stricte (cf. Résultats), aussi seules des horloges moléculaires relâchées ont été utilisées au cours de ces analyses, particulièrement les horloges moléculaires relâchées non corrélées (« uncorrelated relaxed clock »). Dans ces modèles, le taux d'évolution de chaque branche est déterminé à partir d'une distribution sous-jacente exponentielle ou lognormale (Drummond *et al.* 2006).

Certains paramètres ont été fixés pour l'analyse.

Dans le modèle idéal de coalescence, la taille effective de la population étudiée (N_e) est constante et les générations ne sont pas chevauchantes : c'est le modèle Wright – Fisher (Fisher 1930 ; Wright 1931). Les membres de cette population partagent donc la même capacité à diverger que les membres de la population recensée lors de l'étude. Dans la réalité, la taille effective d'une population varie en fonction des générations. Néanmoins,

une population virale est de fait non sexuée, aussi, la taille d'une génération n'est pas dépendante par exemple, du sex-ratio de la précédente. De plus, dans un système épidémique, la population reste globalement stable, la croissance et l'extinction de la population se faisant au gré de la virulence qui s'équilibre elle-même avec la transmission du pathogène. Nous avons donc considéré que la taille de la population de deux générations était constante.

L'intervalle dans lequel est compris le taux d'évolution moléculaire μ, a été fixé selon le taux observé dans les trois alignements étudiés et selon la formule décrite en 2-1-2-3. Afin de ne pas trop contraindre les calculs, la valeur minimale observée a été arbitrairement divisée par un facteur 10 et la valeur maximale multipliée par un facteur 2. La valeur initiale de μ a ensuite été placée sur le minimum observé. Au final, la valeur initiale de μ pour le gène B646L a été fixée à $5,3 \times 10^{-3}$ pour un intervalle de $[5,3 \times 10^{-4} – 2,8 \times 10^{-1}]$, à $7,5 \times 10^{-3}$ et un intervalle de $[7,5 \times 10^{-4} – 3,3 \times 10^{-1}]$ pour le gène E183L et $5,36 \times 10^{-3}$ pour un intervalle de $[5,36 \times 10^{-4} – 1,99 \times 10^{-1}]$ pour le gène CP204L.

Une autre valeur initiale de μ a été testée : lors de l'analyse testant l'hypothèse de l'horloge moléculaire stricte en maximum de vraisemblance (logiciel PAML), un taux de substitution global pour l'ensemble de l'arbre a été calculé. Ce résultat a été utilisé comme valeur initiale de μ avec un intervalle $[0 – +\infty]$.

Les MCMCMC ont été tournées pendant 10^8 générations avec un échantillonnage des arbres tous les 10^4 arbres générés. Et, comme dans l'analyse MCMC effectuée avec le logiciel Mr Bayes, un burn-in de 25% a été réalisé, soit les 2500 premières itérations des chaines de Markov générées ont été écartées des suites de l'analyse. A titre de comparaison, toutes les analyses ont été effectuées en duplicat. L'arbre consensus généré par le logiciel a lui aussi été construit après un burn-in de 25% des arbres générés.

Les arbres finalisés ont alors été visualisés et annotés à l'aide du logiciel FigTree, développé par A. Rambaud (http://tree.bio.ed.ac.uk/software/figtree/).

RÉSULTATS

1- Abondement de la base de données dédiée au virus PPA avec les séquences malgaches

1-1- Isolement des souches de virus PPA malgaches

Le virus PPA se cultive sur macrophages alvéolaires de porcs, car le virus possède un tropisme particulier pour ces cellules (Minguez *et al.* 1988 ; Wardley *et al.* 1977). Le virus débute sa réplication sitôt son entrée, mais l'effet cytopathique (ECP) induit n'est visible en général qu'à partir du troisième jour après le début de l'infection. Il se caractérise d'abord par une granulosité accrue du cytoplasme des cellules, qui vont tendre à se vacuoliser avec l'avancement de l'infection. Les macrophages se regroupent par petits nombres (5 – 6 cellules) jusqu'à parfois former des cellules multinuclées. Bien que la souche malgache du virus PPA ait été décrite comme non cytopathique (Gonzague *et al.* 2001), nous avons cependant pu suivre l'évolution de l'infection jusqu'à destruction quasi intégrale du tapis cellulaire, soit pendant 7 jours.

Un autre moyen de suivre l'intensité et la progression de l'infection des macrophages alvéolaires, outre la lecture des ECP, est de procéder au test de l'hémadsorption. En effet, le génome de certains virus PPA contient un gène qui code pour une protéine capable de se lier au récepteur CD2 cellulaire, le gène EP402R. Ainsi, ajouter des globules rouges frais de porcs dans le milieu de culture après l'infection permet la fixation des globules rouges à la membrane des macrophages et l'observation de « rosettes ». Il est alors aisé de suivre le développement de l'infection. Bien que la souche malgache du virus PPA ait été déclarée non hémadsorbante (Gonzague *et al.* 2001), nous avons constaté une proportion identique de virus hémadsorbants (HAD+) et non hémadsorbants (HAD-) parmi nos isolats. Sur les 21 souches que nous avons isolées, 9 étaient HAD+ et 10 étaient HAD-, deux souches n'ayant pas été testées.

1-2- Production des séquences d'intérêt

L'amplification des 4 gènes cibles au moyen des couples d'amorces spécifiques a été réalisée avec succès et les produits d'amplification générés ont été séquencés après avoir

été clonés dans le plasmide pCR2®.1-topo. Les fragments séquencés ont montré une taille de 478 pb pour le gène B646L, 766 pb pour le gène E183L et 743 pb pour le gène CP204.

2- Analyse approfondie de la phylogénie du virus PPA

2-1- Analyse des alignements

2-1-1- Vérification des alignements

L'amplification de la partie c – terminale du gène B646L qui code pour la protéine majeure de la capside virale a permis de générer un fragment de 478 pb à partir des souches de virus malgaches. Les alignements multiples des 12 premières séquences produites ont montré une homologie de 100% entre ces isolats, ainsi qu'avec la première souche malgache isolée dont la séquence du gène B646L a été utilisée lors de la première étude phylogénétique du virus PPA (Bastos *et al.* 2003).

Les alignements multiples incluant toutes les séquences disponibles que nous avons intégrées à notre base de données ont été réalisés avec 361 séquences d'isolats en provenance d'Afrique, d'Europe, de la Fédération de Russie, d'Amérique du Sud et des Caraïbes. Certaines de ces séquences contenaient des erreurs de séquençage, telles que des bases indéterminées (indiquées par des « N » dans les séquences) ou encore la présence de codons stop à l'intérieur des séquences. La présence de tels codons ne peut être biologiquement valide. En effet, un codon stop signifie l'arrêt de la traduction en protéine de l'ARN messager transcrit. Il en résulte alors une molécule tronquée, dont la conformation ne peut en général plus assurer le rôle biologique qui lui est assigné. Or, la protéine VP72, codée par le gène B646L, est une protéine de structure fondamentale, utilisée lors de la morphogénèse du virus et permettant l'encapsidation de l'ADN viral dans le virion. Toute modification substantielle de sa conformation sera donc vraisemblablement létale pour le virus, puisque de nouveaux virions ne pourront être assemblés. Nous avons donc retiré des alignements les séquences erronées. De même, l'ensemble des trous dans les alignements ont été retirés, ainsi que le codon stop terminant les séquences, et qui n'offre pas d'intérêt phylogénétique. Une vérification identique a été réalisée pour chacun des alignements des gènes E183L et CP204L.

2-1-2- Pertinence du signal phylogénétique contenu dans les alignements

L'analyse de la saturation des substitutions par le logiciel DAMBE version 5.2.0.14 permet de savoir si les alignements utilisés sont pertinents pour effectuer des analyses

phylogénétiques. En effet, un taux de substitution trop élevé entre les séquences étudiées empêche les outils de phylogénie moléculaire de construire l'arbre « vrai », c'est-à-dire l'arbre le plus en adéquation avec les données (Xia *et al.* 2003). En effet, si les séquences ont subi une saturation totale des substitutions, la similarité entre les séquences dépendra exclusivement des fréquences nucléotidiques observées à chaque site, ce qui ne reflète pas les relations phylogénétiques entre les isolats viraux. Pour renforcer les résultats de l'analyse, deux arbres opposés en termes de topologie sont testés : un arbre symétrique et un arbre asymétrique (très improbable).

Le logiciel DAMBE ne permet pas l'analyse d'alignements dans lesquels se trouvent des séquences identiques. Les analyses ont donc été réalisées sur des jeux de séquences uniques, soit 67 séquences pour le gène B646L, 72 séquences pour le gène E183L et 65 séquences pour le gène CP204L.

Les résultats de cette analyse ont montré pour chacun des trois alignements un index de saturation des substitutions (I_{ss}) très inférieur à l'index de saturation des substitutions critique ($I_{ss.c}$), et avec une P_{valeur} égale à 0 (Tableau 1). L'$I_{ss.c}$ est défini en fonction de la longueur critique de l'arbre, c'est-à-dire de la longueur de l'arbre dont la probabilité qu'il soit le plus en adéquation avec le jeu de données dont il est l'expression est égale à 0,95. Ainsi, un I_{ss} significativement inférieur à l'$I_{ss.c}$ permet d'affirmer que le phénomène de saturation des substitutions ne s'applique pas aux alignements des séquences étudiées, séquences qui sont donc pertinentes en termes de signal phylogénétique.

Gène	n taxa	I_{ss}	$I_{ss.c}$Sym.	P_{valeur}	$I_{ss.c}$Asym.	P_{valeur}
B646L	32	0,133	0,691	0,00	0,362	0,00
E183L	32	0,115	0,702	0,00	0,377	0,00
CP204L	32	0,081	0,702	0,00	0,377	0,00

Tableau 1 : Analyse de la saturation des substitutions pour les gènes B646L, E183L et CP204L. n taxa est le nombre de taxa testés, le phénomène de saturation des substitutions n'étant plus un problème au-delà de 16 pour les topologies symétrique et 32 pour les topologies asymétriques.

2-1-3- Détection des recombinaisons

Le logiciel RDP3 version 3 n'a détecté aucune recombinaison au sein des alignements des gènes B646L et CP204L. En revanche, de nombreuses recombinaisons ont été détectées dans l'alignement du gène E183L (Figure 11).

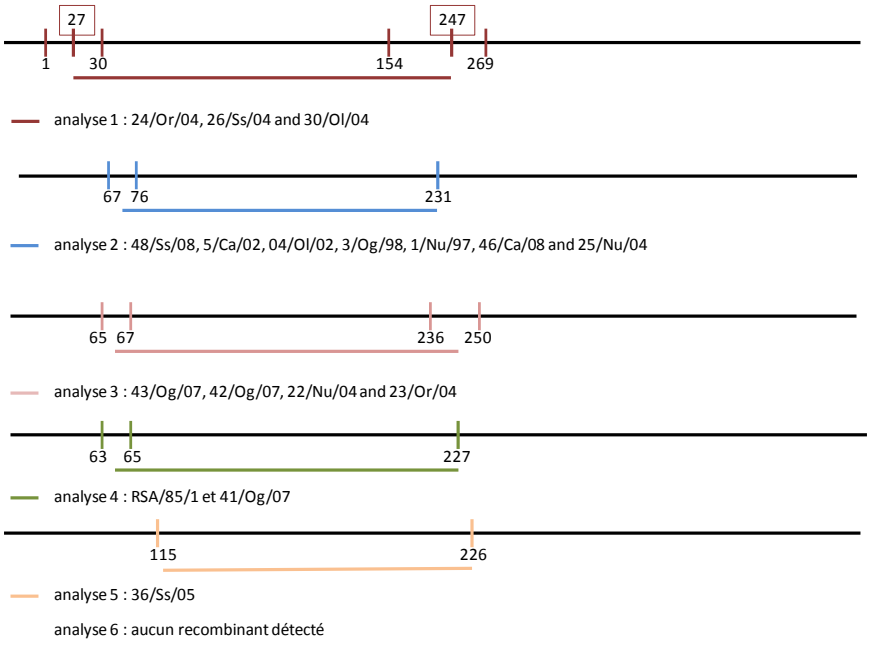


Figure 11 : Récapitulatif des évènements de recombinaison détectés au sein des séquences du gène E183L. Les zones de recombinaison sont indiquées ainsi que les 17 isolats concernés (16 isolats italiens et un isolat Sud Africain). Afin d'éviter le biais évolutif dû à ces recombinaisons, ces isolats ont été retirés du jeu de données.

Lors de la première analyse de détection des recombinants effectuée sur ce dernier alignement, 107 séquences (sur un total de 251) ont montré un signal de recombinaison, avec des points d'entrée et de fin des fragments recombinants variés, mais néanmoins se situant dans les mêmes régions des séquences, c'est-à-dire d'environ 30 à 250 nt. Sur ces 107 séquences, 79 ont montré des signaux évidents de recombinaisons, tandis que seules des traces de recombinaisons (non significatives d'un point de vue statistique) étaient détectées chez les 28 autres. Pour 78 d'entre elles, deux séquences étaient à l'origine de toutes ces recombinaisons, à savoir deux isolats en provenance de Sardaigne : 24/Or/04 (parent mineur, soit la séquence recombinante majoritaire) et 26/Ss/04 (parent majeur). Les analyses faites par le logiciel ont cependant laissé la possibilité que, plutôt que d'avoir 78 séquences recombinantes, les parents mineur et majeur soient en fait les séquences recombinantes. Nous avons donc décidés de retirer ces dernières de l'alignement. La dernière séquence, sarde, l'isolat 30/Ol/04, avait pour parent majeur un isolat également sarde et un parent majeur indéterminé. Nous avons décidé de retirer également de l'alignement cet isolat, puis nous avons procédé à une nouvelle détection des recombinaisons dans le nouvel alignement produit.

La seconde analyse a détecté 13 séquences recombinées, toutes d'origine sardes, dont le parent mineur était indéterminé et le parent majeur était une souche Sud Africaine, la souche MKUZY/79. Sur ces 13 détections, 6 sont apparues à l'état de traces. Le virus de la PPA est aujourd'hui endémique en Sardaigne, provoquant régulièrement des foyers épidémiques. Par essence, une île est un système relativement clos, et il est peu probable

que les souches de virus circulant puissent avoir recombiné avec une souche d'Afrique du Sud, isolée quelques vingt ans auparavant. En revanche, il serait moins surprenant que ces isolats puissent recombiner entre eux au fil des épisodes épidémiques survenus depuis 1978, date de l'entrée du virus dans l'île. Nous avons donc décidé de retirer de l'alignement les 7 séquences détectées de façon certaine, puis nous avons procédé à une nouvelle analyse avec le logiciel RDP3.

La troisième analyse a détecté 6 séquences de Sardaigne recombinantes dont 2 à l'état de traces. Le parent mineur était indéterminé et le parent majeur une souche namibienne isolée en 1989, la souche SPEC/209. Bien que le logiciel ait laissé la possibilité que cette souche soit le recombinant, nous avons décidé de retirer les 4 souches détectées pour les mêmes raisons que celles exposées ci-dessus, puis recommencé l'analyse.

Lors de cette quatrième détection, au sein de 75 séquences des traces de recombinaisons ont été détectées, et des signaux de recombinaisons évidents dans 9 autres séquences, séquences de provenances diverses. Le parent mineur, RSA/85/1 (Afrique du Sud) et le parent majeur 41/Og/07 (Sardaigne), pouvant être eux-mêmes les recombinants. Les parents mineur et majeur ont donc été retirés de l'alignement.

La cinquième analyse n'a mis en évidence qu'une seule séquence recombinée, 36/Ss/05, toujours de Sardaigne, qui a été retirée de l'alignement.

Enfin, au cours de la sixième analyse, aucun évènement de recombinaison n'a pu être détecté.

Au final, se sont 17 séquences qui ont été retirées de l'alignement originel du gène E183L (16 séquences provenant d'isolats sardes et 1 séquence d'un isolat d'Afrique du Sud) pour effectuer les analyses phylogénétiques ultérieures. La saturation des substitutions a à nouveau été testée en utilisant l'alignement final avec un I_{ss} toujours très inférieur à l'$I_{ss.c}$, signe que le signal phylogénétique n'a pas été perdu lors du retrait des séquences recombinantes. Les analyses phylogénétiques suivantes ont donc été réalisées sur des alignements comprenant 351 séquences du gène B646L sur une longueur de 399 pb, 123 séquences du gène CP204L d'une longueur de 543 pb et 234 séquences de 480 pb du gène E183L.

2-1-4- Composition des alignements

L'analyse de la composition des alignements effectuée par le logiciel DnaSP version 5 a montré quelques 110 (27,6%) sites nucléotidiques polymorphiques (c'est-à-dire informatifs en termes de divergence) dans l'alignement du gène B646L, 263 sites (54,7%) dans l'alignement du gène E183L et 154 (28,2%) au sein de l'alignement du gène CP204L. Parmi les 110 sites polymorphiques du gène B646L, 59 se sont avérés être des transversions et 68

des transitions pour un total de 62 substitutions synonymes et 60 substitutions non synonymes. Le nombre moyen de substitutions entre deux séquences (k) étant de 13,629 nucléotides et la divergence par site entre deux séquences (π) égale à 0,03416. Le taux de substitution par site et par an (μ) observé, déterminé manuellement, étant quant à lui compris entre 5,3 x 10^{-3} et 1,4 x 10^{-1} substitutions/site/an.

Pour le gène E183L, se sont 152 transversions et 174 transitions qui ont été observées, pour 88 substitutions silencieuses et 214 substitutions d'acides aminés. Le nombre moyen κ étant égal à 34,960 pour une divergence π par site entre deux séquences de 0,07283. Le taux μ observé étant compris dans l'intervalle [1,15 x 10^{-2} – 2,15 x 10^{-1}] substitutions/site/an.

Enfin, les séquences du gène CP204L ont montré 65 transversions et 113 transitions, pour 83 substitutions synonymes et 88 substitutions non synonymes, avec κ égal à 36,588, π égal à 0,06738 et μ observé compris entre 5,57 x 10^{-3} et 1,02 x 10^{-1} substitutions/site/an.

Il est à noter le nombre élevé de substitutions non synonymes concernant le gène E183L. La protéine p54, pour laquelle il code, étant exposée à l'enveloppe externe du virus, elle est soumise à la pression du système immunitaire de l'hôte, ce qui se traduit par un nombre important de substitutions nucléotidiques dans le gène entrainant un changement d'acide aminé dans la protéine traduite. Cependant, si la protéine VP72 codée par le gène B646L est protégée de cette pression de sélection car située à l'intérieur du virion, la protéine p32, codée par le gène CP204L, est elle aussi intégrée dans l'enveloppe virale externe. Le gène devrait donc montrer davantage de substitutions non synonymes. Or, il s'avère que la protéine p32, qui est la protéine précoce la plus exprimée au début de l'infection, est retrouvée dans le noyau de la cellule infectée. Dans le noyau, la protéine p32 est associée à la protéine hnRNP-K (Hernaez *et al.* 2008), une protéine impliquée dans la régulation de la transcription et de la traduction des gènes (Michelotti *et al.* 1996). La capacité de protéine p32 à interagir avec la protéine hnRNP-K est due à une conformation spécifique c'est-à-dire l'enchainement de certains acides aminés à des sites précis, elle ne peut donc pas subir de mutations trop importantes sans risquer de perdre son activité biologique.

L'ensemble de ces résultats est synthétisé dans le tableau 2, ci-dessous.

Gene	Taille (nt)	Polymorphic sites	Tv	Ts	Syn.	Non Syn.	π	k	d_N/d_S	μ observed
B646L	399	110 (27,6%)	59	68	62	60	0,03416	13,63	0,223	[0,0053 - 0,14]
B646L Sans P°	393									[0,07 - 0,002]
E183L	480	263 (54,7%)	152	174	88	214	0,07283	34,96	0,286	[0,0115 - 0,215]
E183L Sans P°	453	215 (51,2%)								[0,0075-0,166]
CP204L	543	154 (28,2%)	65	113	83	88	0,06738	36,59	1,123	[0,00557 - 0,102]
CP204L Sans P°	534	146 (27,3%)								[0,097 – 0,0054]

Tableau 2 : Composition nucléotidique des alignements. Avec Tv : transversion, Ts, transition, Syn. et non Syn. : substitution(s) muette(s) (syn.) ou ayant entrainé des mutations dans la protéine traduite (non syn.). π : diversité moyenne entre deux séquences (nombre moyen de nucléotides différents par site entre deux séquences), k : nombre moyen de nucléotides différents entre deux séquences et μ observé : le taux de substitution par site et par an observé. Tv : transversion, Ts : transition, Sans P° : sans les codons soumis à pression de sélection.

2-2- Reconstructions phylogénétiques

2-2-1- Enracinement des arbres

L'analyse de l'alignement des séquences complètes d'acides aminés du gène de la protéine majoritaire de capside (MCP) du virus PPA avec la MCP de virus des familles virales les plus proches des *Asfarviridae* par le logiciel TREEFINDER a fourni deux modèles évolutifs : le modèle LG + I et le modèle betHIV + I (se référer à l'annexe 4 pour l'expression complète des modèles précisant les taux de substitutions, les fréquences nucléotidiques, la valeur de la distribution gamma (α) et la fraction des sites invariables (Θ)).

Deux arbres ont ainsi été construits selon ces modèles. Dans le cas qui nous occupe, c'est-à-dire la connaissance du point de divergence entre les groupes externes *Poxviridae*, *Iridoviridae* et *Phycodnaviridae* et les séquences du groupe d'intérêt *Asfarviridae*, nous n'avons pas tenté de trouver l'arbre phylogénétique le plus « vrai ». En effet, l'application des deux modèles évolutifs à l'alignement des séquences protéiques a montré que les groupes externes et le groupe des virus PPA se rejoignaient sur la branche menant aux génotypes décrits VIII et IX-X (Figure 12). Ce sera donc sur cette branche que la racine sera placée lors de la construction des arbres phylogénétiques décrivant les relations entre les isolats de virus PPA.

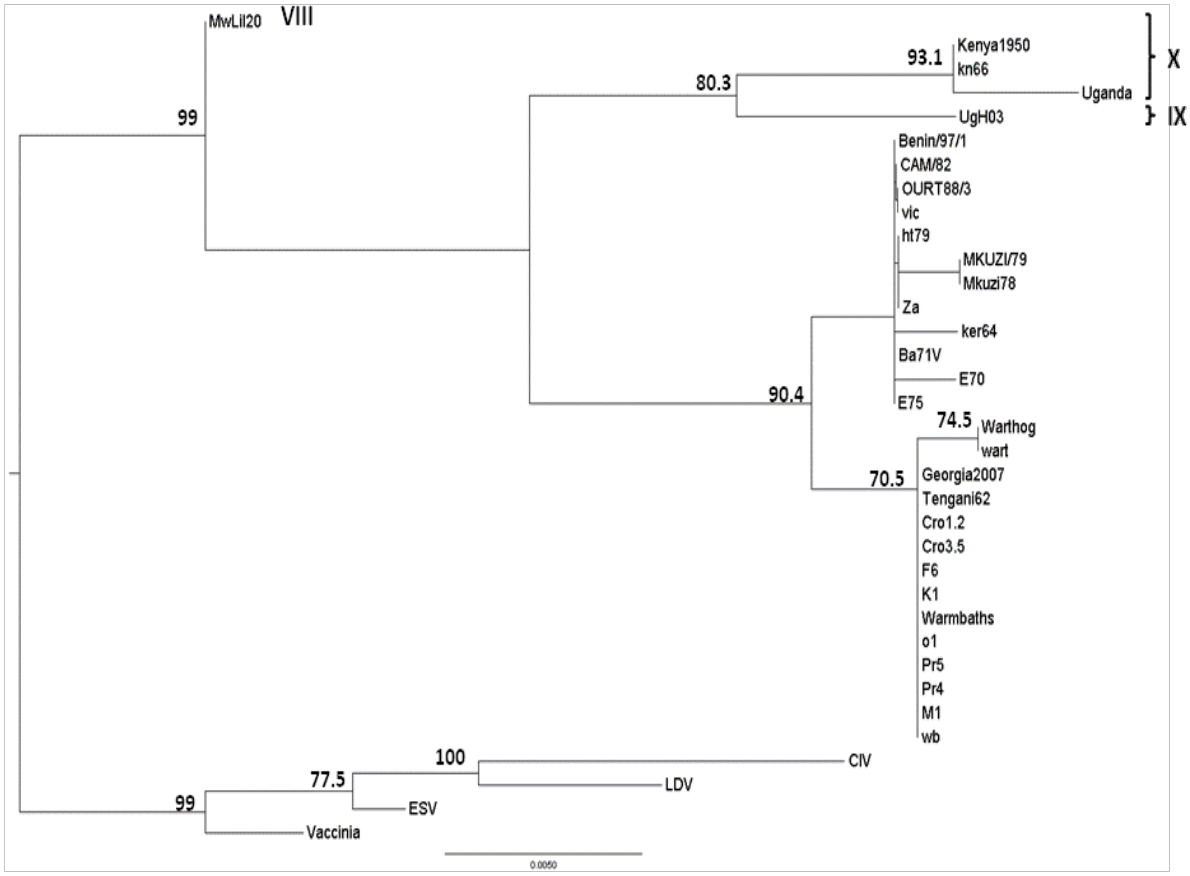


Figure 12 : Arbre phylogénétique en maximum de vraisemblance construit selon le modèle évolutif LG + I par le logiciel TREEFINDER avec la méthode de bootstrap ELW. Il est issu d'un alignement des séquences protéiques du gène de la protéine majeure de la capside (MCP). Le groupe contenant les virus CIV et LDV (*Iridoviridae*), le virus ESF (*Phycodnaviridae*) et le virus de la vaccine (*Poxviridae*) représente le groupe externe. Les autres isolats viraux représentés appartiennent aux *Asfarviridae*. Les virus PPA et le groupe externe divergent sur la branche menant au génotype décrit VIII et aux génotypes IX et X. C'est le point d'enracinement des isolats de virus PPA.

2-2-2- Reconstructions phylogénétiques utilisant le gène B646L

2-2-2-1- Maximum de vraisemblance

Les modèles évolutifs proposés par le logiciel TREEFINDER et considérés comme les plus en adéquation avec l'alignement généré avec les séquences de la partie c – terminale du gène B646L ont été les suivants : sous le critère d'information AIC, le modèle évolutif a été le modèle TN93 + $\Gamma 5$ (Tamura & Nei 1993), un modèle à six paramètres. L'analyse sous les critères d'information AICc et BIC, a fourni un modèle évolutif identique à cinq paramètres : HKY85 + $\Gamma 5$ (Hasegawa *et al.* 1985) (se référer à l'annexe 4 pour l'expression complète des modèles).

Les deux arbres ont donc été construits avec ces deux modèles à l'aide du logiciel TREEFINDER et du ré-échantillonnage ELW (1000 bootstraps). Un arbre phylogénétique a

également été construit selon le modèle GTR (modèle le plus complexe à 9 paramètres). La comparaison des arbres a permis de sélectionner l'arbre phylogénétique le plus vrai, c'est-à-dire le plus en adéquation avec le jeu de données de séquences étudié. L'arbre sélectionné a été celui construit avec le modèle évolutif HKY85 + Γ5 (Figure 13). Il montre l'existence de quatre branches principales : la première branche regroupe les génotypes I, II, XVII et XVIII, la seconde les génotypes III, IV, V, VI, VII, XIX, XX, XXI et XXII, ainsi qu'un isolat non encore génotypé, Cro3.5. La troisième branche réunit les génotypes VIII, XI, XII, XIII, XV et XVI, et replace l'isolat TAN/08/MAZIMBU, précédemment classé en tant que génotype XV comme feuille unique d'une branche proche du génotype VIII. Enfin, la quatrième branche réunit les génotypes IX et X, en provenance de la région des grands lacs africains, berceau de la Peste porcine africaine, où le virus suis un cycle selvatique ancien incluant les suidés sauvages et les tiques molles *Ornithodoros*.

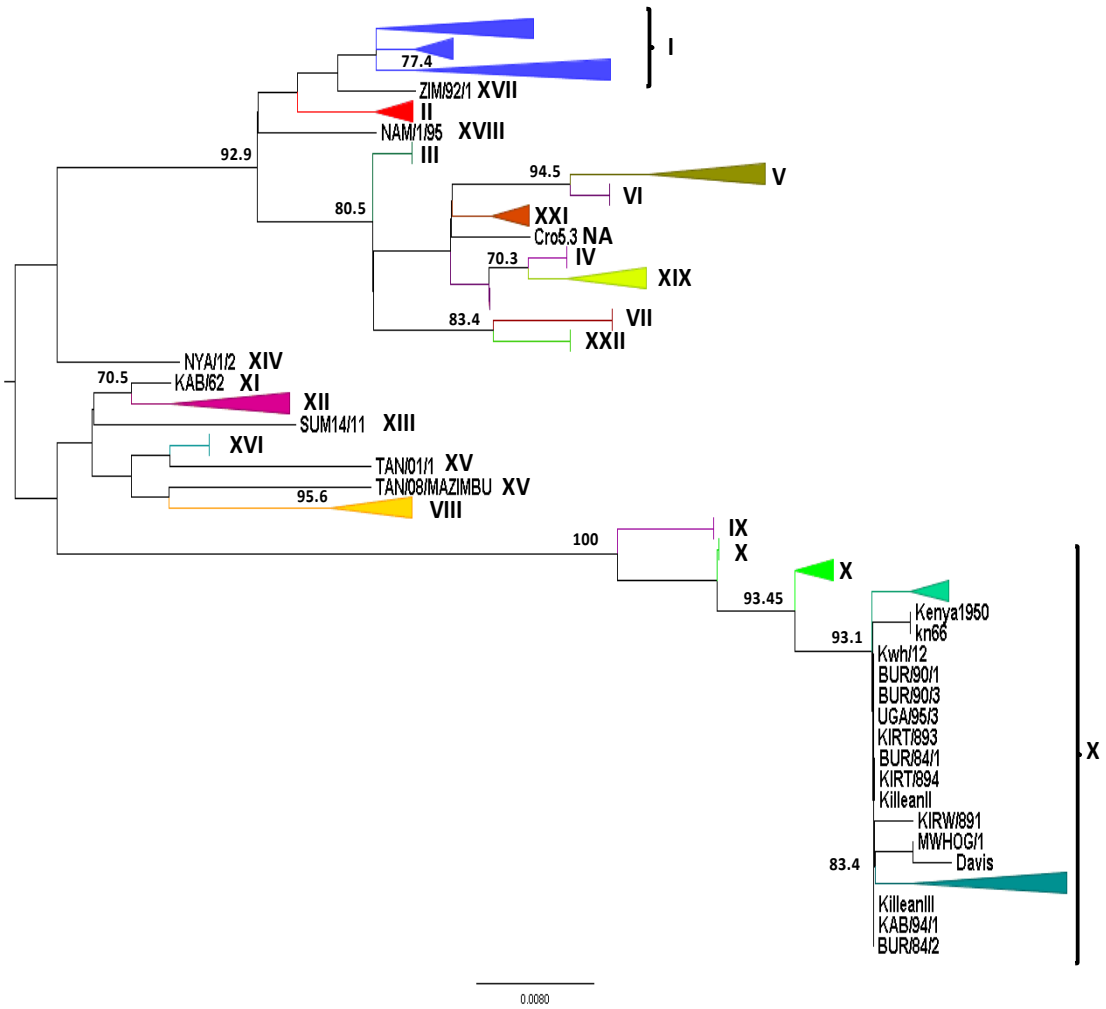


Figure 13 : Arbre phylogénétique du gène B646L construit en maximum de vraisemblance avec le modèle évolutif HKY85 + Γ5. Les isolats ont été classés selon la classification en génotypes établie par Bastos *et al*. (2003), Lubisi *et al*. (2005) et Boshoff *et al*. (2006). L'isolat Cro3.5 d'Afrique du Sud, jusqu'ici non encore classé

(NA : non assigné) est la feuille unique d'une branche de l'arbre. De même, l'isolat tanzanien TAN/08/MAZIMBU, classé au sein du génotype XV lors d'une étude précédente (Misinzo *et al.* 2011) ne se ségrége pas avec l'isolat TAN/01/1 (génotype XV), mais se rapproche du génotype VIII.

Pour des raisons de lisibilité, les isolats de chaque génotype ont été condensés. Ainsi, l'arbre phylogénétique ne montre aucune polytomie, la construction de l'arbre sous un modèle évolutif en bifurcation semblant alors idéale pour résoudre les relations entre les isolats du jeu de données de séquences du gène B646L étudié. Cependant, la ségrégation des isolats à l'intérieur du génotype I et surtout du génotype X montre de nombreuses multifurcations (Figure 14).

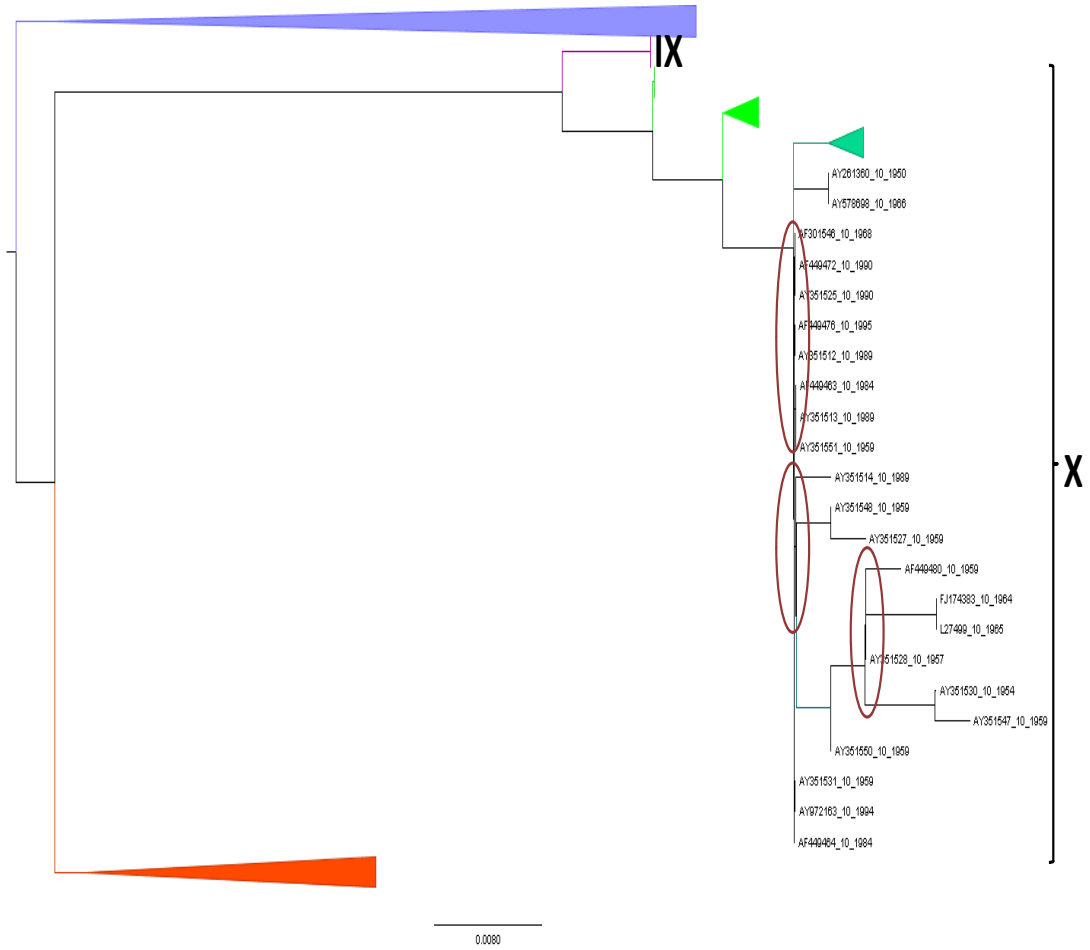


Figure 14 : Détails de la discrimination des isolats du génotype X lors de la construction de l'arbre phylogénétique utilisant les séquences du gène B646L avec le modèle évolutif HKY85 + Γ5 en maximum de vraisemblance. Les polytomies sont indiquées par des cercles rouges. Ces multifurcations indiquent que, si la méthode de construction de l'arbre par bifurcations valide la discrimination en génotypes, elle ne permet pas de résoudre finement les relations entre isolats au niveau intra-groupe.

2-2-2-2- Inférence bayésienne

Deux arbres phylogénétiques ont été construits avec l'alignement des séquences du gène B646L au moyen des chaines de Markov avec technique de Monte Carlo. Le premier a été généré en suivant le modèle évolutif HKY85 + $\Gamma 5$ sélectionné après l'analyse en maximum de vraisemblance. L'analyse a été stoppée alors que les MCMC avaient été tournées pendant 4 x 10^6 itérations, pour un LRT (comparaison statistique *a posteriori* des arbres générés) oscillant de façon stable autour de 0,04 sur 4 x 10^5 générations, soit 10% des arbres générés. Le second arbre phylogénétique par inférence bayésienne a été construit selon le modèle GTR + $\Gamma 5$. Les MCMC on été tournées également durant 4 x 10^6 itérations et stoppées lorsque la valeur des LRT s'est stabilisée au cours des 4 x 10^5 dernières itérations. Le LRT lors de l'arrêt de l'analyse était égal à 0,064.

Dans les deux cas, l'arbre phylogénétique consensus a été construit après avoir rejeté les 25 premiers pourcents des arbres générés, afin que les probabilités *a priori* initiales, qui sont générées aléatoirement, et donc reflétant moins la réalité des relations entre isolats, n'interfèrent pas dans l'arbre final. L'échantillonnage des arbres au cours des MCMC s'effectuant chaque 10^3 arbres, l'arbre consensus a donc été construit à partir des 3000 derniers arbres générés. La comparaison entre les arbres construits selon les deux modèles évolutifs HKY85 + $\Gamma 5$ et GTR + $\Gamma 5$ a été réalisée manuellement, selon la puissance discriminative des deux modèles. L'arbre le plus discriminant a été déterminé comme étant celui généré avec le modèle HKY85 + $\Gamma 5$ (Figure 15).

En comparaison de l'arbre phylogénétique construit en maximum de vraisemblance et un modèle évolutif identique (Figure 13), l'on peut constater un pouvoir discriminant inférieur pour l'analyse en inférence bayésienne. En effet, seule la branche comprenant les génotypes IX et X est construite à l'identique. La branche contenant les génotypes VIII, XI, XII, XIII, XV et XVI présente une polytomie (génotype XIII) qui n'existait pas dans l'arbre en maximum de vraisemblance. L'isolat TAN/08/MAZIMBU, placé dans la littérature au sein du génotype XV mais discriminé de ce génotype lors de l'analyse en maximum de vraisemblance en se rapprochant du génotype VIII, se ségrége à nouveau avec l'isolat TAN/01/2. La valeur de bootstrap affectée au nœud qui sépare TAN/08/MAZIMBU et TAN/02/1, est suffisamment élevée (98) pour que ces deux isolats ne puissent êtres considérés comme un seul et même génotype. C'est surtout parmi les autres génotypes que la discrimination est nettement moins précise : si les génotypes I et II forment toujours des groupes bien différenciés, ils ne forment plus une branche majeure avec les génotypes XVII et XVIII. L'arbre ne présente que trois branches principales, celles contenant les génotypes IX et X ainsi que celle incluant les génotypes VIII, XI, XII, XIII, XV et XVI restent identiques, tandis que les autres génotypes ne se différencient plus en, d'une part les génotypes I, II, XVII et XVIII, et d'autre part les génotypes III, IV, V, VI, VII, XIX, XX, XXI et XXII, mais sont regroupés au sein d'une troisième branche principale. La discrimination des génotypes III, IV, V, VI, XIX, XX, XXI et XXII s'est avérée difficile à résoudre, l'arbre montrant de nombreuses polytomies, et les

génotypes III, XIX et XX n'étant pas discriminés en clusters. L'emploi de cet arbre pour les analyses en datation moléculaire ne sera donc pas pertinent.

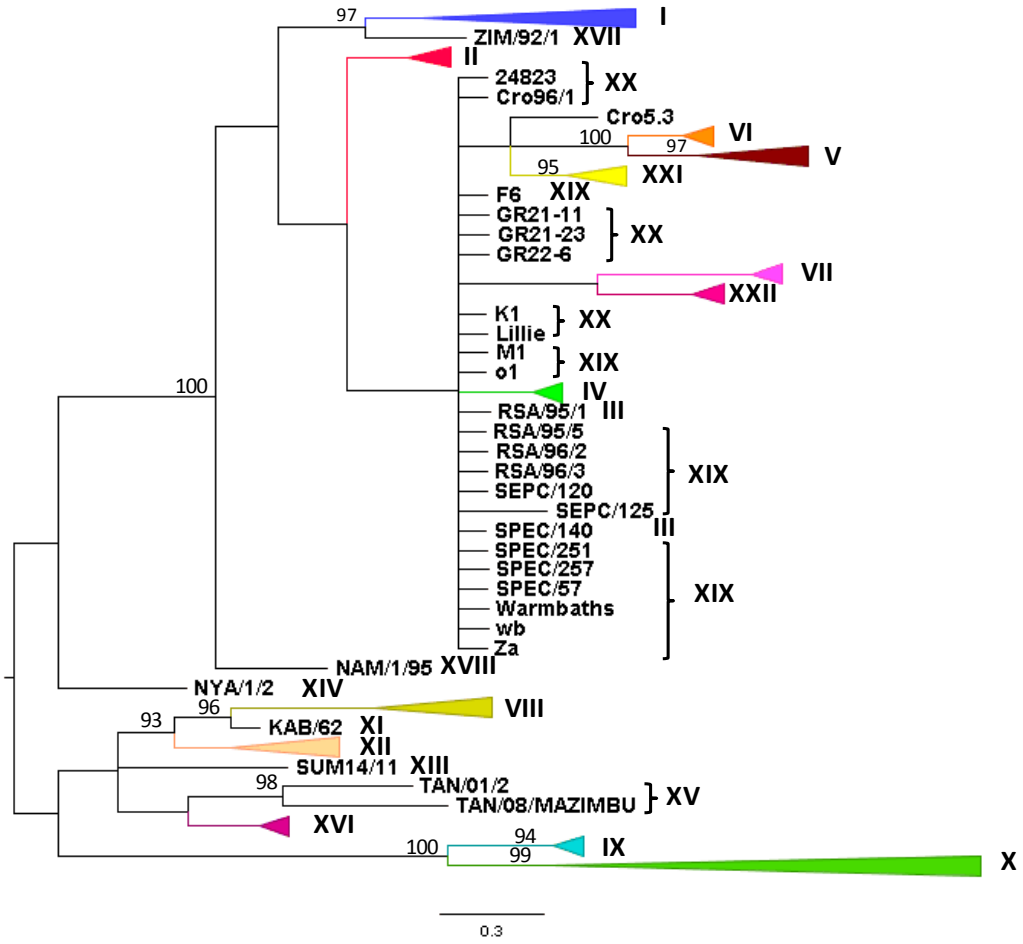


Figure 15 : Arbre phylogénétique construit par inférence bayésienne à l'aide du logiciel Mr Bayes sous le modèle évolutif HKY85 + Γ5. Les isolats on été classés selon la classification en génotypes décrite dans la littérature par Bastos *et al*. (2003), Lubisi *et al*. (2005) et Boshoff *et al*. (2006). Seuls les valeurs de bootstrap supérieures à 90 sont indiquées.

2-2-3- Classification des isolats de virus PPA

2-2-3-1- Classification par la méthode des distances

La première approche que nous avons utilisée afin de classer les isolats de virus PPA a été celle utilisant la matrice des distances générée lors de la construction de l'arbre phylogénétique à partir de l'alignement des séquences du gène B646L. La matrice choisie a été celle générée avec l'arbre HKY85 + Γ5. Nous avons choisi d'établir la classification à partir du gène B646L car ce sont les séquences de ce gène qui ont servi de socle à la classification des souches de virus initiée en 2003 par Bastos *et al*.

Dans cette matrice de distance, les isolats ont été regroupés selon les génotypes décrits dans la littérature, c'est-à-dire de I à XXII. Les isolats non encore génotypés dans la littérature ont été implémentés dans les génotypes décrits en fonction de leur position dans l'arbre phylogénétique induit. Deux isolats ont été placés à part : le premier, Cro3.5, n'avait pas été encore génotypé et constitue une feuille unique terminant une branche de l'arbre, tandis que le second, TAN/08/MAZIMBU avait été placé au sein du génotype XV alors que l'arbre phylogénétique que nous avons construit l'en différencie. La moyenne des distances intra et intergénotypes, la distance minimale entre les groupes ainsi que la distance maximale à l'intérieur d'un même groupes ont alors été calculées (Tableau 3).

	Moyenne intra-groupes	Maximum intra-groupe	Moyenne intergroupes	Minimum intergroupes
Distance	0,00228	0,0262	0,0538	0,0051

Tableau 3 : Moyennes des distances intra et intergroupes, minimum intergroupes et maximum intra-groupe. Ces valeurs on été déterminées en groupant les isolats en fonction des génotypes décrits dans la littérature (I à XXII).

Le calcul de ces valeurs montre que, si la moyenne des distances intergroupes est 20 fois plus élevée que la moyenne des distances intra-groupes, le maximum de distance au niveau intra-groupe est plus de 50 fois plus élevé que le minimum intergroupes.

Pour comprendre la répartition des distances entre isolats, la distribution des distances a été déterminée à partir de cette matrice de distance. Pour ce faire, les fréquences de ces distances ont été estimées. Dans le cas idéal, les distances entre isolats au niveau intra-groupe et au niveau intergroupes seraient séparées de façon absolue. Ainsi, la courbe des fréquences représenterait une double courbe de type gaussienne, la limite entre ces deux courbes passant par 0 (Figure 16).

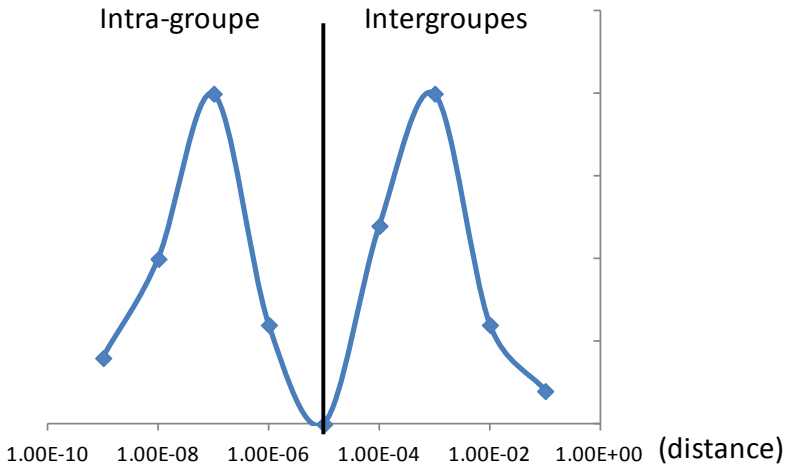


Figure 16 : Distribution idéale des distances au sein d'un jeu de séquences : la courbe passe par un minimum égal à 0 qui correspond à la délimitation absolue des différents clusters.

La réalité biologique ne produit pas une distribution aussi claire des distances intra et intergroupes. La figure 17 montre le graphique des fréquences estimées, et donc la distribution des distances entre les séquences du gène B646L. La matrice d'intervalles utilisée couvre l'ensemble des distances calculées par le logiciel TREEFINDER, 10^{-15} à 2 x 10^{-1}, la distance maximale entre deux séquences du gène B646L étant égale à 1, 19 x 10^{-1}. Les séquences identiques ne présentant pas d'intérêt en termes de distance, elles ne sont pas représentées sur le graphique. La courbe des fréquences ne montre pas de façon absolue la limite entre les distances intra et intergroupes. Une extrapolation a donc été réalisée en projetant les courbes intra et intergroupes sur l'axe des abscisses. La zone comprise entre ces deux projections devant représenter les séquences pour lesquelles il est difficile de statuer.

A partir de cette répartition des distances entre les séquences, les isolats viraux groupés selon les génotypes décrits et montrant des distances intra-groupes supérieures au minimum intergroupes (c'est-à-dire appartenant à la zone « douteuse ») ont été replacés en sous-groupes. Le maximum intra-groupe et le minimum intergroupes ont alors été recalculés selon cette nouvelle classification. Ainsi, de proche en proche, l'ensemble des isolats a été reclassé. Or, il est apparu que cette méthode de classification a produit de nombreux groupes ne comportant qu'un seul isolat, ce qui, sur le plan biologique, comme à l'observation de l'arbre phylogénétique, n'est pas pertinent. Certains groupes de séquences sont particulièrement le siège de telles difficultés. Ce sont les séquences appartenant aux groupes décrits comme étant les génotypes I et X, qui montrent une grande hétérogénéité entre certains isolats.

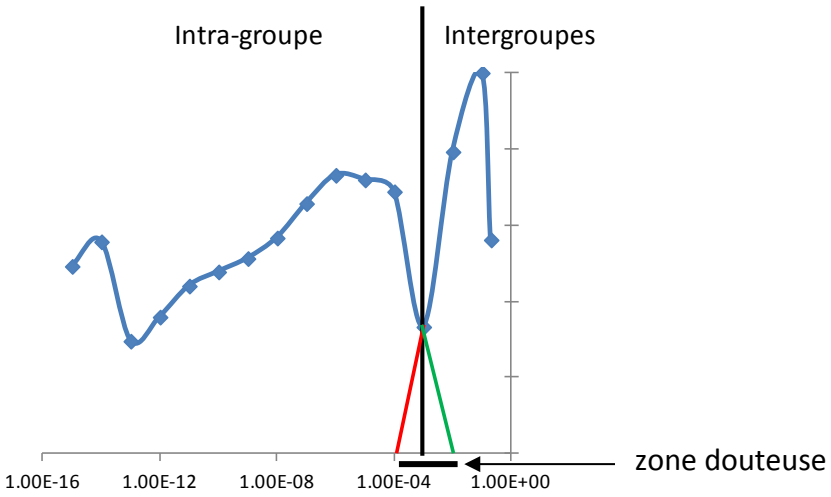


Figure 17 : Distribution des distances entre les séquences du gène B646L des isolats du virus PPA. La courbe de distribution ne passe pas par un minimum égal à 0 et ne produit donc pas une définition absolue des distances au niveau intra et intergroupes. La projection sur l'axe des abscisses montre la zone dans laquelle les isolats sont difficilement classables.

Il a donc été établi que la classification des isolats en fonction des distances entre les séquences géniques introduisait un biais conduisant à trop individualiser les isolats. Ainsi, si les distances reflètent une réelle différence entre les séquences, la clustérisation des isolats viraux ne peut reposer sur elles seules, et elles ne peuvent donc être le seul support de la classification, tout du moins selon la nomenclature décrite dans la littérature. Le constat d'une telle disparité entre les isolats au niveau intra-groupe pose la question des relations réelles qui unissent les virus PPA entre eux. De plus, la présence de polytomies au sein des branches principales de l'arbre phylogénétique indique que la phylogénie par bifurcation pourrait ne pas être optimale pour la compréhension de ces relations.

2-2-3-2- Analyse en réseau

Afin d'envisager les liens entre isolats viraux prenant en compte les polytomies, ou plutôt leur donnant un sens, et ainsi mieux comprendre les relations qui unissent les isolats de virus PPA et en établir une classification claire, des analyses en réseau ont été réalisées.

2-2-3-2-1- Détermination d'un réseau d'haplotypes par le logiciel TCS

Ce logiciel permet de construire un arbre phylogénétique dans lequel sont indiquées le nombre de substitutions existant entre deux séquences où groupes de séquences. Ainsi, le chemin évolutif peut être déterminé non seulement en termes de variation nucléotidique, mais aussi en termes de liens de parenté entre les séquences. Si la distance entre les séquences n'a pas permis de conceptualiser finement les différents groupes de virus PPA, le nombre de substitutions permettant de passer d'un groupe à l'autre pourrait fournir une indication claire de la limite stricte entre deux de ces groupes. De plus, la détermination du réseau d'haplotypes ne se limite pas à dénombrer les mutations entre des séquences ou groupes de séquences. En effet, un tel réseau permet d'établir des connexions multiples entre les séquences, c'est-à-dire d'exprimer comment une séquence peut avoir été générée à partir non pas d'une, mais de plusieurs séquences préexistantes.

La figure 18 montre le réseau d'haplotypes construit à partir de l'alignement des séquences du gène B646L du virus PPA par le logiciel TCS. Le réseau met en exergue le groupe de séquences le plus représenté, soient les isolats appartenant au génotype I. Ce groupe est symbolisé par l'isolat sarde 16/Og/04. C'est à partir de cette séquence de référence que le réseau s'organise, en ajoutant progressivement les substitutions permettant de passer d'une séquence à une autre. En revanche, contrairement à la représentation en arbre phylogénétique par bifurcation, que construisent les méthodes de maximum de vraisemblance et d'inférence bayésienne, plusieurs chemins permettent d'expliquer cette transition d'une séquence à une autre. Par exemple, à partir de la séquence de référence 16/Og/04, deux chemins évolutifs permettent de passer au groupe suivant, le groupe des isolats malgaches, symbolisé par l'isolat Ambaton01. Le premier trajet compte cinq substitutions : la substitution d'un G en A en position 75 aboutit à l'isolat zambien ZAM/01/1. De là, deux substitutions (A – T en position 75 et T – C en position 6) permettent de rejoindre un nœud intermédiaire. De ce nœud intermédiaire, deux autres substitutions amènent à l'isolat Ambaton01 (A – G en position 162 et G – A en position 216). L'autre chemin ne comporte que quatre mutations : T – C en position 6 qui mène à un nœud interne, puis T – G en position 75. Cette mutation conduit au nœud interne placé sur le premier trajet, les deux dernières substitutions sont donc les mêmes pour les deux chemins évolutifs. Certains chemins évolutifs sont encore davantage complexes, comme ceux liant les isolats appartenant aux génotypes IV, V, VI, XIX, XX et XXI, ou encore les isolats du génotype X.

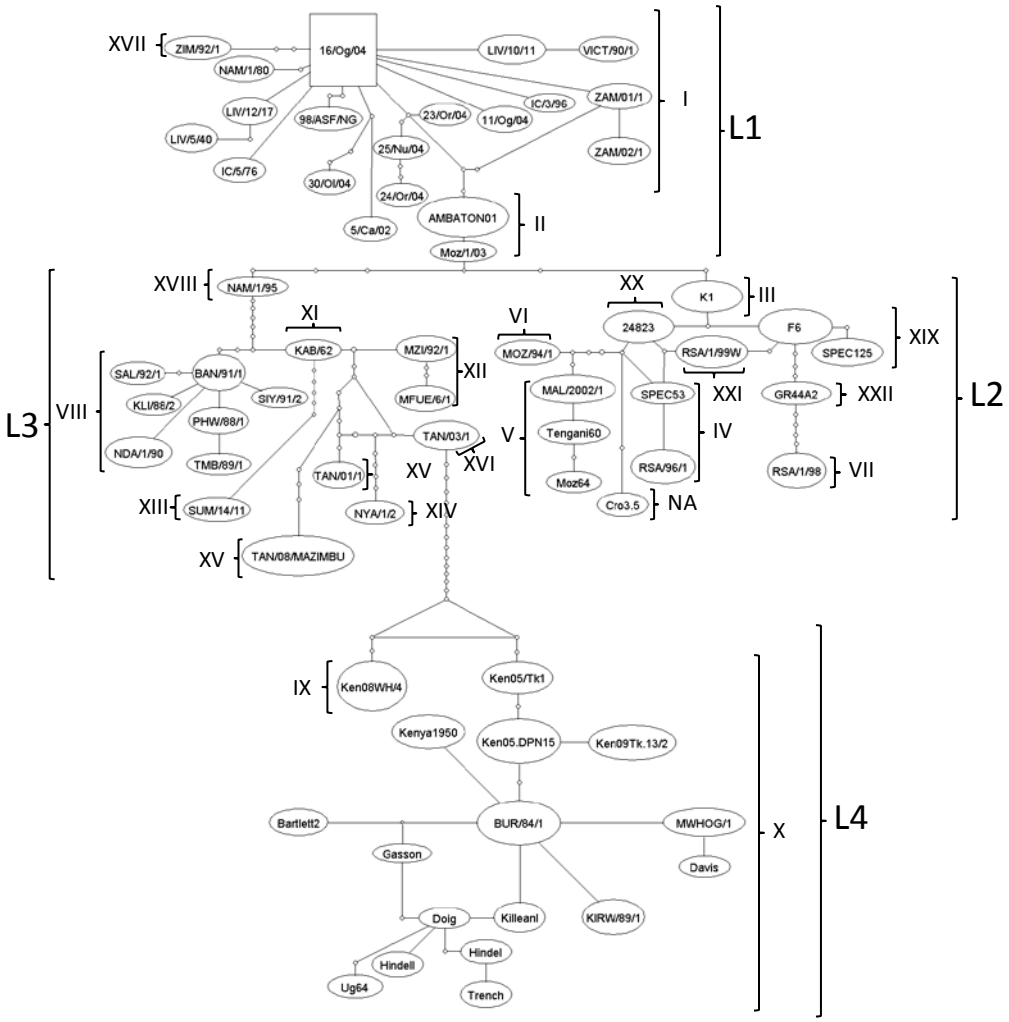


Figure 18 : Réseau d'haplotypes généré par le logiciel TCS à partir de l'alignement des séquences du gène B646L. Les substitutions nécessaires pour passer d'un isolat à un autre sont indiquées par un « o ». Le réseau montre les mêmes quatre lignées principales (L1 à L4) que les arbres construits par bifurcation en maximum de vraisemblance sous le modèle HKY + Γ5. Les relations complexes entre isolats sont clairement indiquées par la présence de plusieurs chemins évolutifs possibles permettant d'expliquer la phylogénèse d'une séquence.

Malgré une représentation plus complexe des relations unissant les isolats ou groupes d'isolats, l'analyse en réseau d'haplotypes fait apparaitre quatre grands groupes d'isolats, correspondant aux grandes branches qui sont montrées par l'analyse en phylogénie par bifurcation, ce qui renforce la ségrégation des isolats et leur regroupement.

2-2-3-2-2- Détermination d'un réseau de partition ou « split-network »

Le réseau de partitions réalisé par le logiciel SPLITSTREE version 4 a également montré la présence de quatre grandes lignées pour résoudre les relations entre isolats de virus PPA (Figure 19). Ces quatre lignées regroupent les mêmes isolats appartenant aux quatre branches principales des arbres phylogénétiques en bifurcation ainsi qu'aux quatre grandes lignées visibles dans le réseau d'haplotypes. De la même façon que le montrait le réseau d'haplotypes, les isolats appartenant au groupe malgache se situent à mi-chemin entre les génotypes I et XVII d'une part, et le groupe contenant les génotypes IV, V, VI, XIX, XX et XXI.

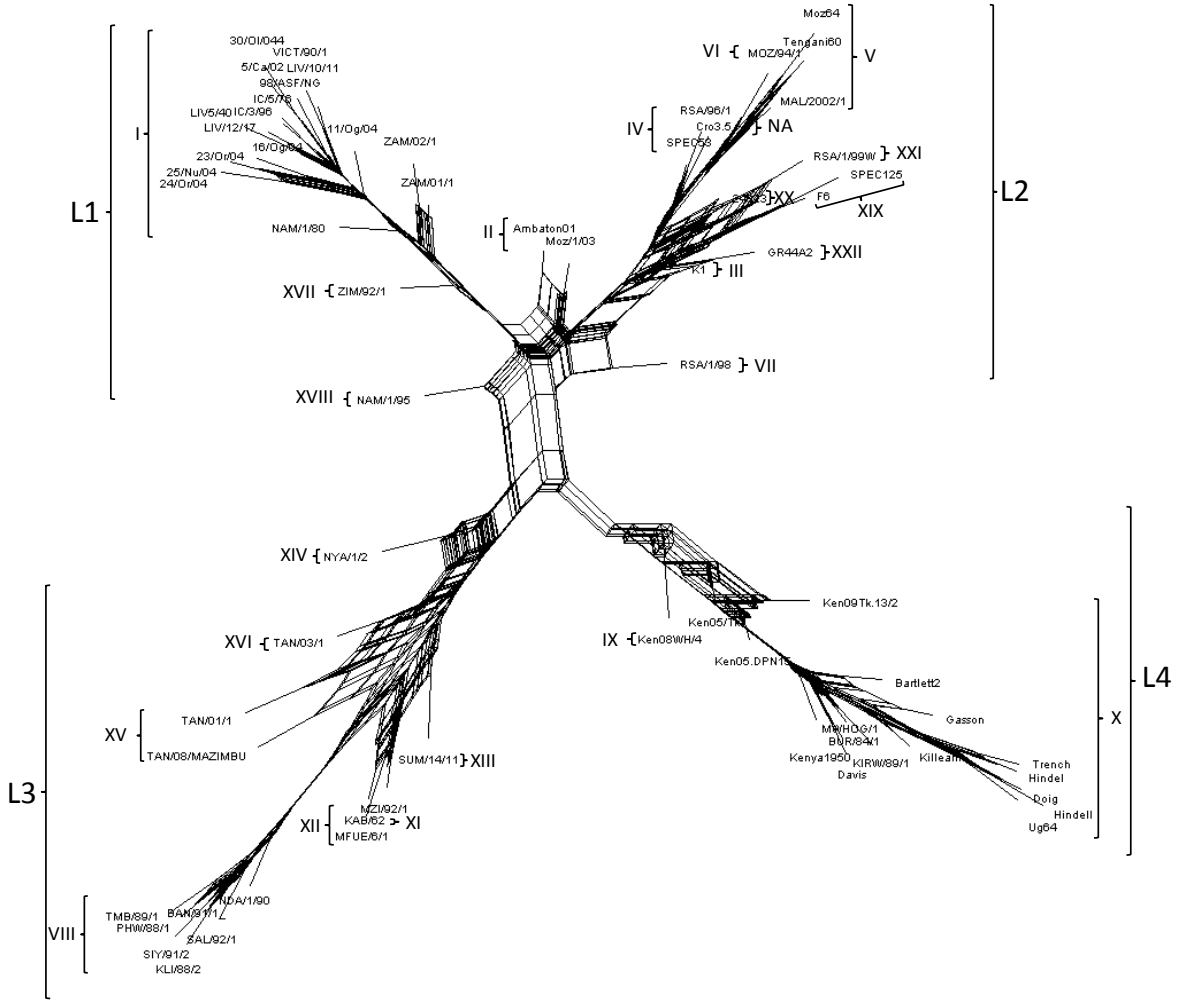


Figure 19 : Réseau de partitions réalisé à l'aide du logiciel SPLITSTREE. Le réseau représenté est issu de 100 ré-échantillonnages (analyse bootstrapée) de l'alignement du gène B646L du virus PPA. Pour des raisons de visibilité, seules les séquences uniques apparaissent dans le réseau. Les quatre principales lignées apparaissent nettement, tout comme lors des analyses précédentes.

2-2-3-3- Détermination de la signature moléculaire des isolats de virus PPA

Les méthodes d'analyses phylogénétiques, que ce soit en bifurcation ou en réseau, ont montré l'existence de quatre lignées distinctes d'isolats de virus PPA. Toutefois, aucune de ces méthodes, pas plus que l'analyse des distances entre les séquences, n'ont pu établir une classification claire des isolats viraux. Afin d'assoir cette classification sur une base biologique, seule réellement à même de grouper des isolats de même origine, nous avons procédé à une analyse de la signature moléculaire des séquences du gène B646L du virus PPA.

L'alignement des séquences du gène B646L a tout d'abord été expurgé de toutes les séquences identiques, pour ne garder que des séquences divergentes. Ainsi, c'est au sein d'un alignement de 67 séquences uniques que la signature moléculaire des géno-groupes a été déterminée. Après avoir retiré les sites nucléotidiques conservés au travers de toutes les séquences, l'analyse de la divergence a porté sur 93 sites d'intérêt. Au total, 35 clusters ont été identifiés par leur signature moléculaire (Figure 20).

En premier lieu, les séquences ont été classées selon les quatre grandes lignées (L1 à L4) déterminées grâce aux arbres phylogénétiques et aux réseaux. Les sites invariants au sein de ses quatre lignées ont à leur tour été retirés des alignements correspondants. Ainsi, la signature moléculaire des lignées, c'est-à-dire la base nucléotidique commune à tous les isolats appartenant à ces lignées, a été déterminée. Par la suite, ces nucléotides communs ont été retirés des alignements afin de laisser apparaitre les sous-lignées. Il est à noter que l'isolat zambien NYA/1/2, décrit comme le génotype XIV dans la littérature, a été classé à part. La branche dont il est la feuille unique se situe, dans l'arbre phylogénétique construit en maximum de vraisemblance, entre les lignées L1 et L3, sans qu'il soit possible de déterminer une parenté plus forte pour l'une ou l'autre de ces lignées. Si l'analyse en réseau de partitions le place au début de la lignée L3, sa signature moléculaire n'en fait cependant pas un membre à part entière.

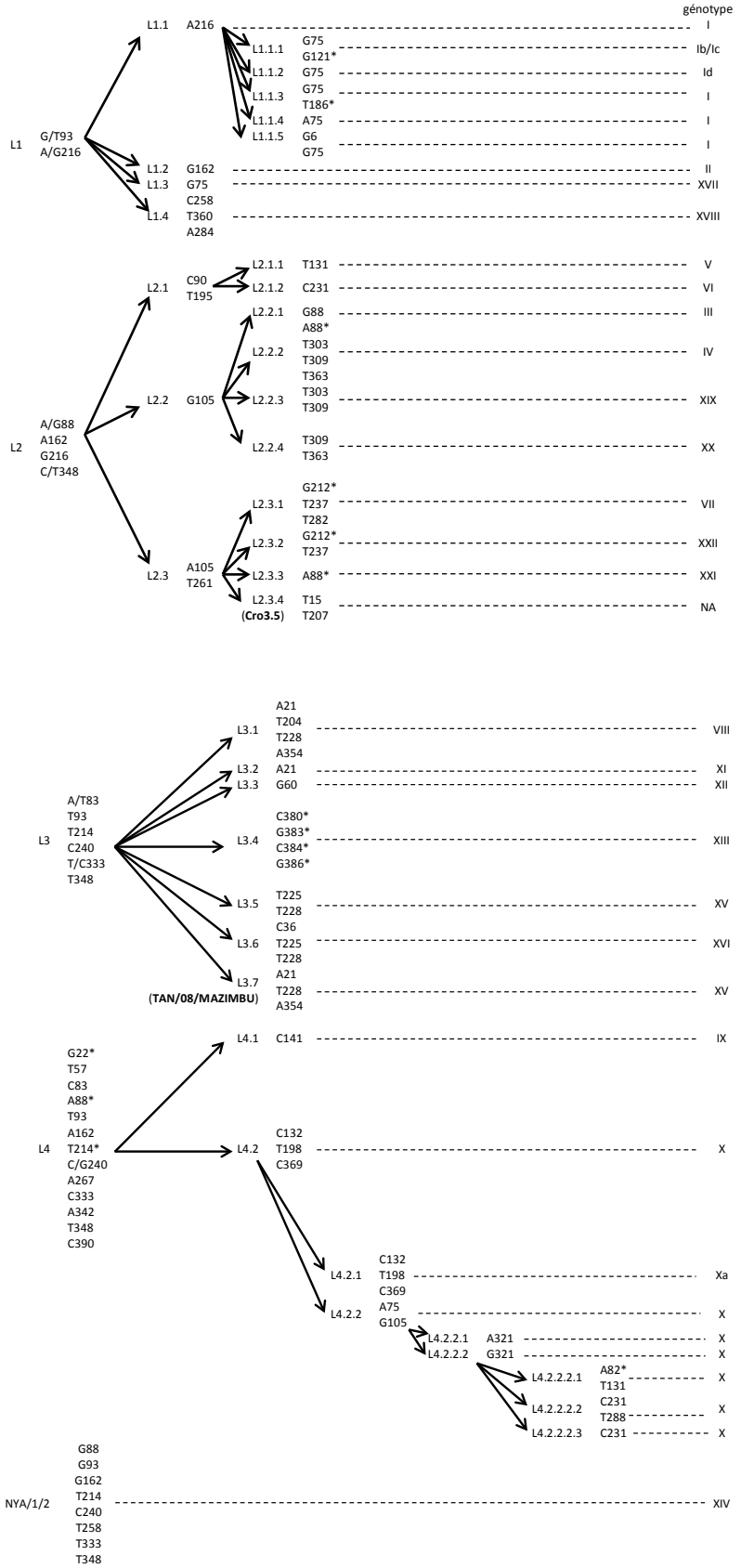


Figure 20 : Signatures moléculaires des différentes lignées et sous-lignées du virus ASFV pour le gène B646L. Les génotypes correspondants sont indiqués dans la colonne de droite. Les substitutions non synonymes sont marquées d'une « * ». 35 clusters ont été identifiés

La lignée L1 compte quatre sous-lignées, qui recouvrent les génotypes décrits I, II, XVII et XVIII. Dont la sous-lignée L1-1 qui se subdivise en cinq sous-sous-lignées. Cette sous-lignée L1-1 correspond au génotype décrit I dans lequel l'arbre phylogénétique montrait une divergence assez forte entre les isolats. Malgré cette divergence élevée entre les isolats, la signature moléculaire de L1-1 est très nette, quoique réduite à un seul nucléotide : une adénine en position 216. Parmi les lignées secondaires de la lignée L1-1, deux d'entre elles sont composées exclusivement d'isolats zambiens ; il s'agit des lignées secondaires L1-1-2 et L1-1-4. La lignée secondaire L1-1-1 est composée d'isolats en provenance de Namibie, d'Afrique du Sud, de Zambie et du Zimbabwe. Enfin, les lignées L1-1-3 et L1-1-5 sont composées d'isolats italiens. La lignée L1-1 correspond historiquement à des isolats européens, ouest africains, caribéens et sud américains. La présence d'isolats en provenance d'Afrique australe formant des lignées secondaires, tout comme la présence d'isolats kenyans évoque une circulation des souches n'ayant pas pour unique origine le cycle selvatique du virus opérant dans ces régions.

La lignée L2 compte trois sous-lignées : L2-1, subdivisée en L2-1-1 (génotype V) et L2-1-2 (génotype 6) ; L2-2, subdivisée en quatre géno-groupes, correspondant aux génotypes III, IV, XIX et XX, et enfin L2-3, subdivisée elle aussi en quatre géno-groupes (génotypes VII, XXI et XXII ainsi que l'isolat Cro3.5, non encore classé).

La lignée L3 compte sept sous-lignées, recouvrant les génotypes VIII, XI, XIII, XV et XVI, ainsi que l'isolat tanzanien TAN/08/MAZIMBU. Ce dernier avait été décrit comme appartenant au génotype XV (Misinzo *et al.* 2010), mais sa signature moléculaire le discrimine clairement de ce géno-groupe.

Enfin, la lignée L4 présente deux sous-lignées avec les génotypes IX (L4-1) et X (L4-2). La sous-lignée L4-2 est celle qui montre le plus de diversité parmi toutes. Elle est subdivisée en deux sous-sous-lignées, L4-2-1 et L4-2-2, elles-mêmes divisées en plusieurs géno-groupes.

Ainsi, l'arbre représenté figure 13, dont les isolats avaient été classés selon les génotypes décrits dans la littérature, peut-il être redéfini selon les lignées que nous avons déterminées en nous basant sur la signature moléculaire des isolats de virus PPA (Figure 21). De même, le réseau d'haplotypes peut-il être représenté selon la nomenclature des lignées que nous avons décrites (Figures 22).

Nous avons également voulu vérifier si cette classification pouvait être consolidée par l'analyse des distances entre les différents clusters. La figure 23 montre la matrice de distance réalisée en clustérisant les séquences selon notre nomenclature. Le seuil de diversité a été placé à 5 fois la moyenne de la diversité entre sous-lignées.

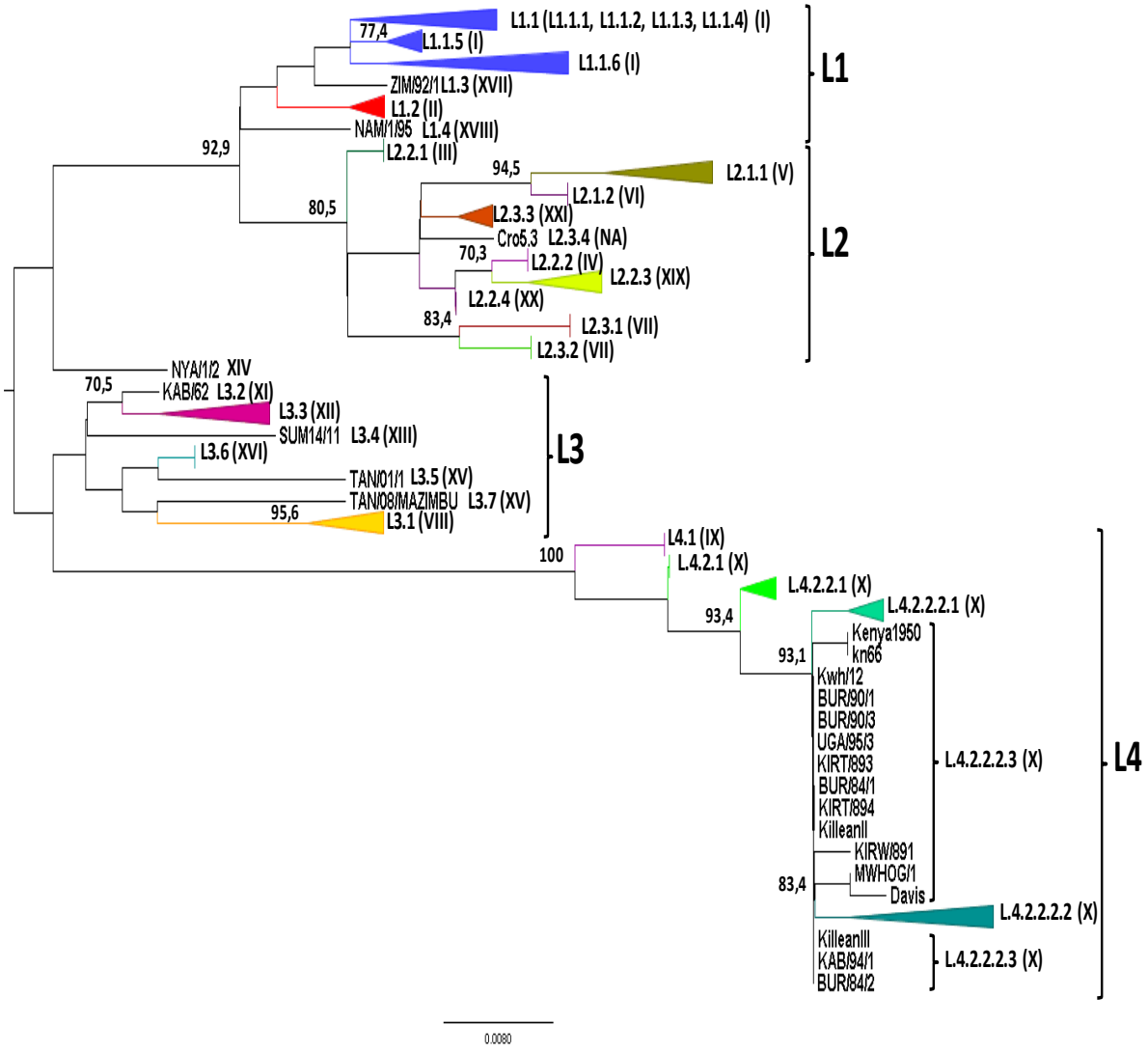


Figure 21 : Arbre phylogénétique classant les différentes lignées de virus PPA, basé sur la signature moléculaire des séquences du gène B646L. Les quatre lignées majeures (L1 à L4) sont clairement définies. L'isolat NYA/1/2 se situe entre les lignées L1-L2 et les lignées L3-L4. La signature moléculaire de cet isolat n'a pas permis de le classer dans l'un ou l'autre de ces groupes. La correspondance avec les génotypes déjà décrits dans la littérature est indiquée entre parenthèse après chaque lignée

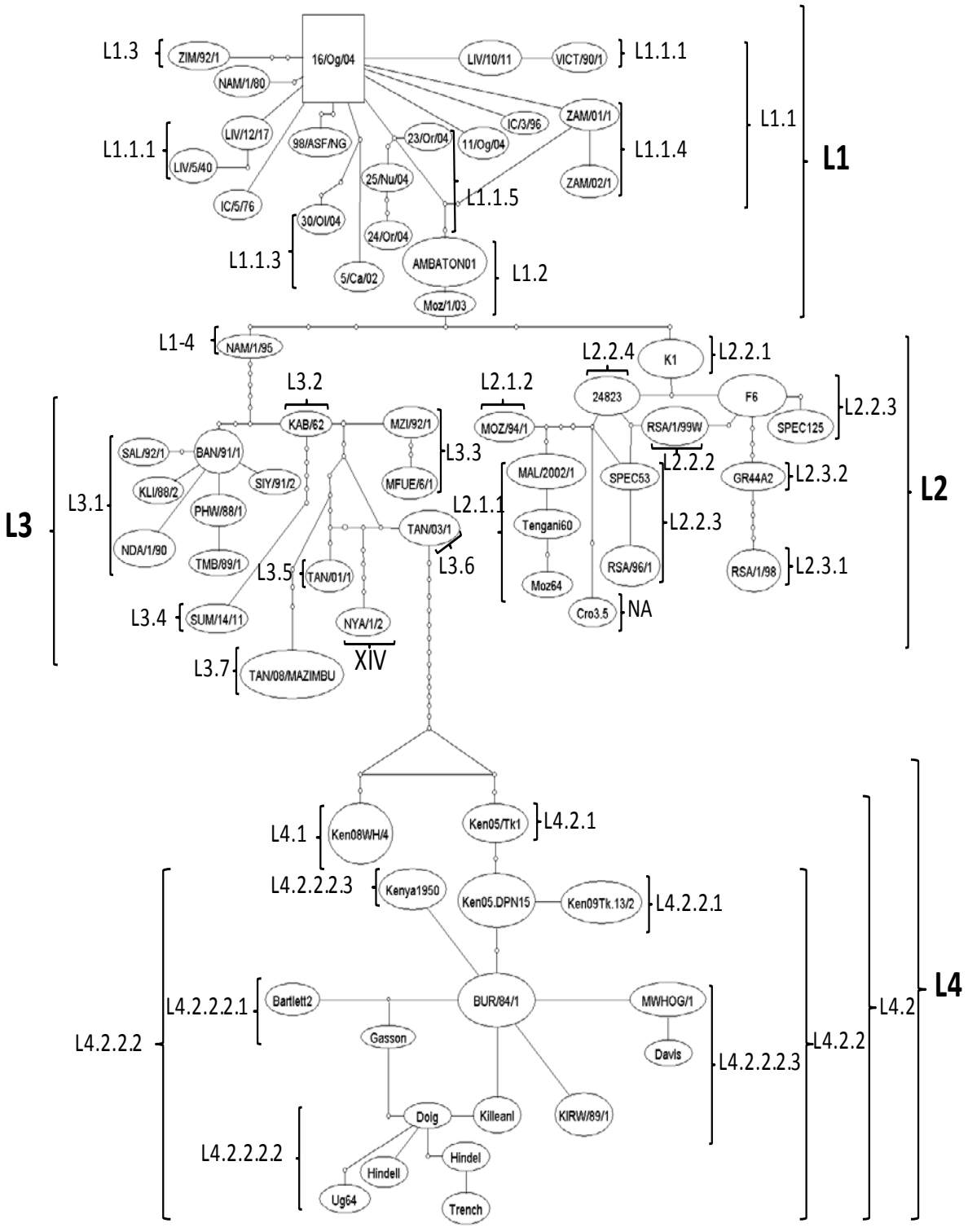
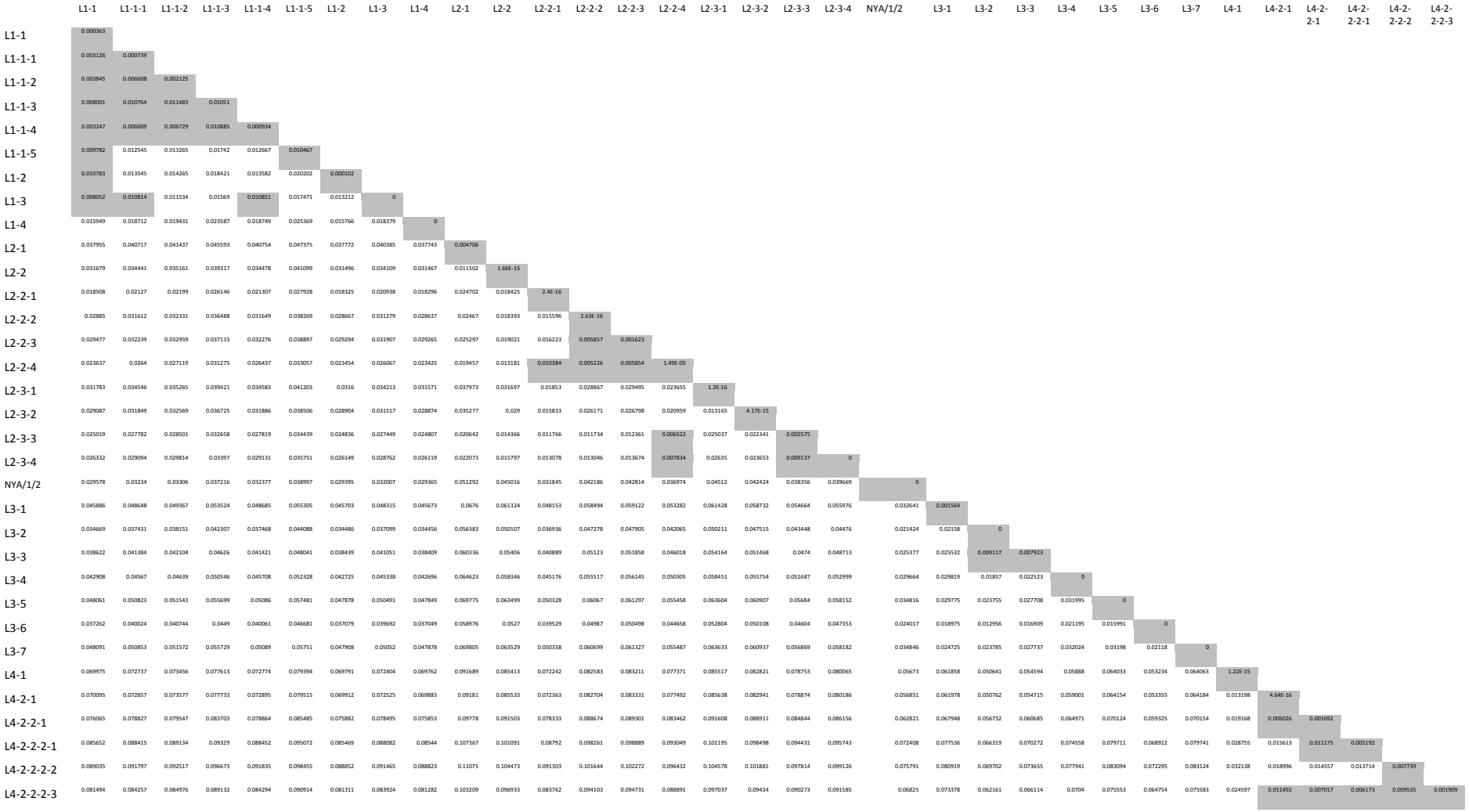


Figure 22 : Réseau d'haplotypes des isolats de virus PPA utilisant le gène B646L. Les différentes lignées que nous avons identifiées ont été replacées.



	L1-1	L1-1-1	L1-1-2	L1-1-3	L1-1-4	L1-1-5	L1-2	L1-3	L1-4	L2-1	L2-2	L2-2-1	L2-2-2	L2-2-3	L2-2-4	L2-3-1	L2-3-2	L2-3-3	L2-3-4	NYA/1/2	L3-1	L3-2	L3-3	L3-4	L3-5	L3-6	L3-7	L4-1	L4-2-1	L4-2-2-1	L4-2-2-2-1	L4-2-2-2-2	L4-2-2-2-3
L1-1	0.000363																																
L1-1-1	0.003126	0.000739																															
L1-1-2	0.003845	0.006608	0.002125																														
L1-1-3	0.008001	0.010764	0.011483	0.01051																													
L1-1-4	0.003247	0.006009	0.006729	0.010885	0.000934																												
L1-1-5	0.009782	0.012545	0.013265	0.01742	0.012667	0.010467																											
L1-2	0.010783	0.013545	0.014265	0.018421	0.013582	0.020202	0.000102																										
L1-3	0.008052	0.010814	0.011534	0.01569	0.010851	0.017471	0.013212	0																									
L1-4	0.015949	0.018712	0.019431	0.023587	0.018749	0.025369	0.015766	0.018379	0																								
L2-1	0.037955	0.040717	0.041437	0.045593	0.040754	0.047375	0.037772	0.040385	0.037743	0.004706																							
L2-2	0.031679	0.034441	0.035161	0.039317	0.034478	0.041099	0.031496	0.034109	0.031467	0.011502	1.66E-15																						
L2-2-1	0.018508	0.02127	0.02199	0.026146	0.021307	0.027928	0.018325	0.020938	0.018296	0.024702	0.018425	2.4E-16																					
L2-2-2	0.02885	0.031612	0.032331	0.036488	0.031649	0.038269	0.028667	0.031279	0.028637	0.02467	0.018393	0.015596	2.63E-16																				
L2-2-3	0.029477	0.032239	0.032959	0.037115	0.032276	0.038897	0.029294	0.031907	0.029265	0.025297	0.019021	0.016223	0.005857	0.001623																			
L2-2-4	0.023637	0.0264	0.027119	0.031275	0.026437	0.033057	0.023454	0.026067	0.023425	0.019457	0.013181	0.010384	0.005226	0.005854	1.49E-05																		
L2-3-1	0.031783	0.034546	0.035265	0.039421	0.034583	0.041203	0.0316	0.034213	0.031571	0.037973	0.031697	0.01853	0.028867	0.029495	0.023655	1.2E-16																	
L2-3-2	0.029087	0.031849	0.032569	0.036725	0.031886	0.038506	0.028904	0.031517	0.028874	0.035277	0.029	0.015833	0.026171	0.026798	0.020959	0.013165	4.17E-15																
L2-3-3	0.025019	0.027782	0.028501	0.032658	0.027819	0.034439	0.024836	0.027449	0.024807	0.020642	0.014366	0.011766	0.011734	0.012361	0.006522	0.025037	0.022341	0.002575															
L2-3-4	0.026332	0.029094	0.029814	0.03397	0.029131	0.035751	0.026149	0.028762	0.026119	0.022073	0.015797	0.013078	0.013046	0.013674	0.007834	0.02635	0.023653	0.009137	0														
NYA/1/2	0.029578	0.03234	0.03306	0.037216	0.032377	0.038997	0.029395	0.032007	0.029365	0.051292	0.045016	0.031845	0.042186	0.042814	0.036974	0.04512	0.042424	0.038356	0.039669	0													
L3-1	0.045886	0.048648	0.049367	0.053524	0.048685	0.055305	0.045703	0.048315	0.045673	0.0676	0.061324	0.048153	0.058494	0.059122	0.053282	0.061428	0.058732	0.054664	0.055976	0.032641	0.001564												
L3-2	0.034669	0.037431	0.038151	0.042307	0.037468	0.044088	0.034486	0.037099	0.034456	0.056383	0.050107	0.036936	0.047278	0.047905	0.042065	0.050211	0.047515	0.043448	0.04476	0.021424	0.02158	0											
L3-3	0.038622	0.041384	0.042104	0.04626	0.041421	0.048041	0.038439	0.041051	0.038409	0.060336	0.05406	0.040889	0.05123	0.051858	0.046018	0.054164	0.051468	0.0474	0.048713	0.025377	0.025532	0.009117	0.007923										
L3-4	0.042908	0.04567	0.04639	0.050546	0.045708	0.052328	0.042725	0.045338	0.042696	0.064623	0.058346	0.045176	0.055517	0.056145	0.050305	0.058451	0.055754	0.051687	0.052999	0.029664	0.029819	0.01857	0.022523	0									
L3-5	0.048061	0.050823	0.051543	0.055699	0.05086	0.057481	0.047878	0.050491	0.047849	0.069775	0.063499	0.050328	0.06067	0.061297	0.055458	0.063604	0.060907	0.05684	0.058152	0.034816	0.029775	0.023755	0.027708	0.031995	0								
L3-6	0.037262	0.040024	0.040744	0.0449	0.040061	0.046681	0.037079	0.039692	0.037049	0.058976	0.0527	0.039529	0.04987	0.050498	0.044658	0.052804	0.050108	0.04604	0.047353	0.024017	0.018975	0.012956	0.016909	0.021195	0.015991	0							
L3-7	0.048091	0.050853	0.051572	0.055729	0.05089	0.05751	0.047908	0.05052	0.047878	0.069805	0.063529	0.050358	0.060699	0.061327	0.055487	0.063633	0.060937	0.056869	0.058182	0.034846	0.024725	0.023785	0.027737	0.032024	0.03198	0.02118	0						
L4-1	0.069975	0.072737	0.073456	0.077613	0.072774	0.079394	0.069791	0.072404	0.069762	0.091689	0.085413	0.072242	0.082583	0.083211	0.077371	0.085517	0.082821	0.078753	0.080065	0.05673	0.061858	0.050641	0.054594	0.05888	0.064033	0.053234	0.064063	1.22E-15					
L4-2-1	0.070095	0.072857	0.073577	0.077733	0.072895	0.079515	0.069912	0.072525	0.069883	0.09181	0.085533	0.072363	0.082704	0.083331	0.077492	0.085638	0.082941	0.078874	0.080186	0.056851	0.061978	0.050762	0.054715	0.059001	0.064154	0.053355	0.064184	0.013198	4.64E-16				
L4-2-2-1	0.076065	0.078827	0.079547	0.083703	0.078864	0.085485	0.075882	0.078495	0.075853	0.09778	0.091503	0.078333	0.088674	0.089301	0.083462	0.091608	0.088911	0.084844	0.086156	0.062821	0.067948	0.056732	0.060685	0.064971	0.070124	0.059325	0.070154	0.019168	0.006026	0.001692			
L4-2-2-2-1	0.085652	0.088415	0.089134	0.09329	0.088452	0.095072	0.085469	0.088082	0.08544	0.107367	0.101091	0.08792	0.098261	0.098889	0.093049	0.101195	0.098498	0.094431	0.095743	0.072408	0.077536	0.066319	0.070272	0.074558	0.079711	0.068912	0.079741	0.028755	0.015613	0.011175	0.005192		
L4-2-2-2-2	0.089035	0.091797	0.092517	0.096673	0.091835	0.098455	0.088852	0.091465	0.088823	0.11075	0.104473	0.091303	0.101644	0.102272	0.096432	0.104578	0.101881	0.097814	0.099126	0.075791	0.080919	0.069702	0.073655	0.077941	0.083094	0.072295	0.083124	0.032138	0.018996	0.014557	0.013714	0.007739	
L4-2-2-2-3	0.081494	0.084257	0.084976	0.089132	0.084294	0.090914	0.081311	0.083924	0.081282	0.103209	0.096933	0.083762	0.094103	0.094731	0.088891	0.097037	0.09434	0.090273	0.091585	0.06825	0.073378	0.062161	0.066114	0.0704	0.075553	0.064754	0.075583	0.024597	0.011455	0.007017	0.006173	0.009535	0.001909

Figure 23 : Distances évolutives entre les différentes lignées et sous-lignées caractérisées d'un point de vue moléculaire. La moyenne intra lignée était de 0,0023, tandis que la moyenne des distances interlignées était de 0,055. Dans cette matrice, la diversité inférieure à 5 X (moyenne diversité intra lignée) est indiquée en gris. L'ensemble des sous-lignées diffèrent les unes des autres par une diversité plus élevée, hormis pour quelques sous-lignées à l'intérieur de L1.1, L1.2, L1.3, L2.2, L2.3 et L4.2.2.

2-2-4- Reconstructions phylogénétiques utilisant le gène E183L

2-2-4-1- Maximum de vraisemblance

Les trois critères d'informations utilisés pour déterminer le meilleur modèle évolutif en maximum de vraisemblance pour expliquer le jeu de données de séquences du gène E183L ont proposé deux modèles différents : les critères d'information AIC et BIC ont proposé le modèle HKY + $\Gamma 5$, avec de légères variations en termes de substitutions, de fréquences nucléotidiques et de la valeur α de la distribution gamma, tandis que le critère d'information AICc proposait le modèle HKY (annexe 4). Les trois arbres phylogénétiques correspondant à ces modèles ont été construits et un support statistique a été calculé pour chaque nœud de l'arbre après les 1000 ré-échantillonnages effectués lors de l'analyse ELW par le logiciel TREEFINDER. L'arbre sélectionné a été celui construit avec le modèle HKY + $\Gamma 5$ proposé par le critère d'information BIC (Figure 24). Si les lignées L1, L3 et L4, telles que définies après l'analyse du gène B646L, sont bien différenciées, la lignée L2 se scinde en deux parties : d'une part les sous-lignées L2.2.1, L2.2.3, L2.3.1 et L2.3.2, et d'autre part les sous-lignées L2.1.1, L2.1.2, L2.2.2, L2.2.3, L2.2.4. Deux isolats mozambicains appartenant à la lignée L1.2 dans la classification B646L se ségrégent

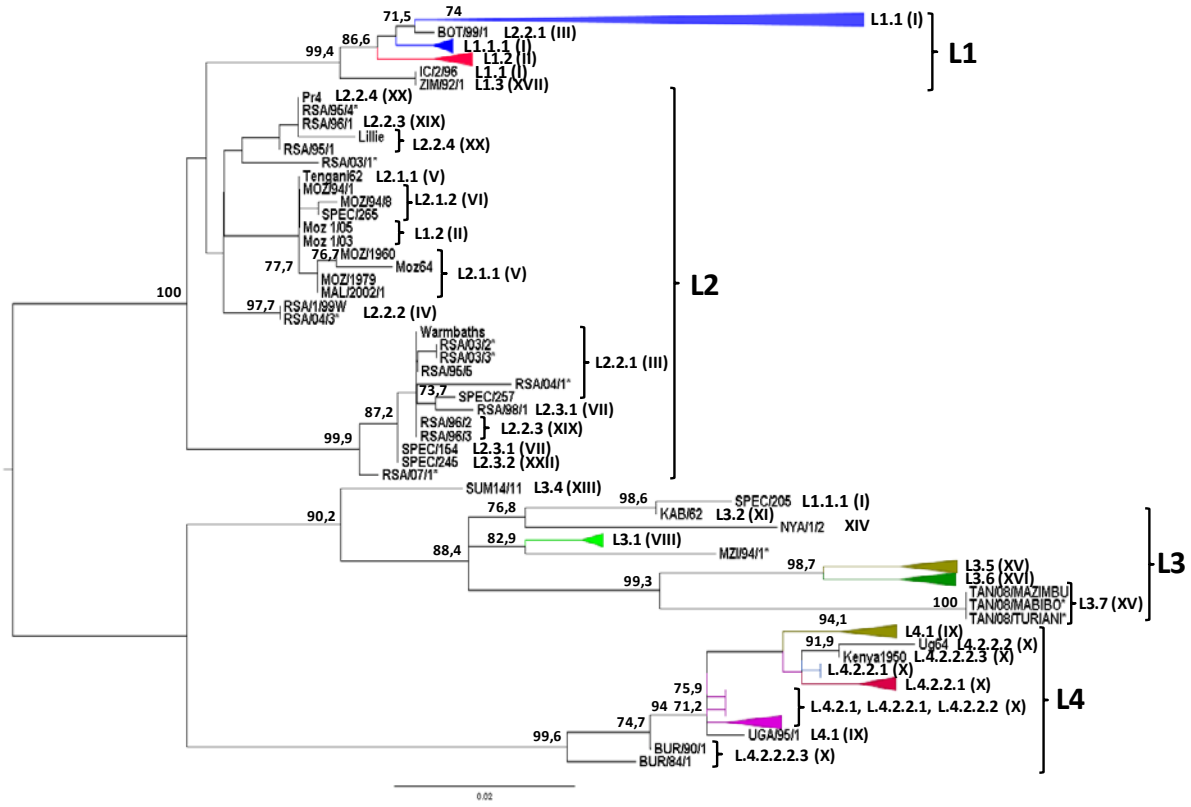


Figure 24 : Arbre phylogénétique construit en maximum de vraisemblance avec le modèle HKY85 + Γ5 à partir des séquences du gène E183L. Les lignées et sous-lignées déterminées lors de l'analyse du gène B646L ont été replacées. Seules les valeurs de bootstraps supérieures à 70 sont indiquées. Les isolats marqués d'un astérisque n'entraient pas dans l'analyse du gène B646L.

2-2-4-2- Inférence bayésienne

Deux arbres phylogénétiques ont été construits avec l'alignement des séquences du gène E183L au moyen des chaines de Markov avec technique de Monte Carlo. Le premier suivant le modèle évolutif HKY85 + Γ5 (sélectionné après l'analyse en maximum de vraisemblance) et le second selon le modèle GTR + Γ5. Avec le modèle HKY85 + Γ5 les MCMC ont été tournée pendant 2 x 10^6 génération. L'analyse sous le modèle HKY85 + Γ5 a été stoppée alors que le LRT, égal à 0,06, n'avait pas atteint la valeur seuil de 0,01, mais était resté stable autour de 0,06 durant 4 x 10^5 générations. Concernant l'analyse selon le modèle GTR + Γ5, les MCMC ont effectué 2.9 x 10^6 pour un LRT final égal à 0.07. L'arbre consensus a été établi, dans les deux cas, après rejet des 25 premiers pourcents des arbres générés, par la compilation de 1500 arbres avec le modèle HKY85 + Γ5 et 2148 arbres avec le modèle GTR + Γ5. L'arbre le plus discriminant a été déterminé manuellement, il s'est agit de l'arbre construit selon le modèle HKY85 + Γ5 (Figure 25).

La comparaison avec l'arbre construit en maximum de vraisemblance selon le même modèle évolutif, l'arbre E183L en inférence bayésienne montre lui aussi une scission de la

lignée L2 qui se ségrége pour une part avec la lignée L1, l'autre part, formant au nouveau une lignée différenciée. On note également une moins bonne discrimination des isolats, principalement en ce qui concerne la lignée L4, pour laquelle la méthode ne parvient pas à différencier en clusters les sous-lignées provenant de la lignée L4.2. De même, l'isolat UGA/95/1 (L4.1, génotype IX) se ségrége-t-il avec les sous-lignées L4.2.1, L4.2.2.1 et L4.2.2.2.3, elles-mêmes non différenciées.

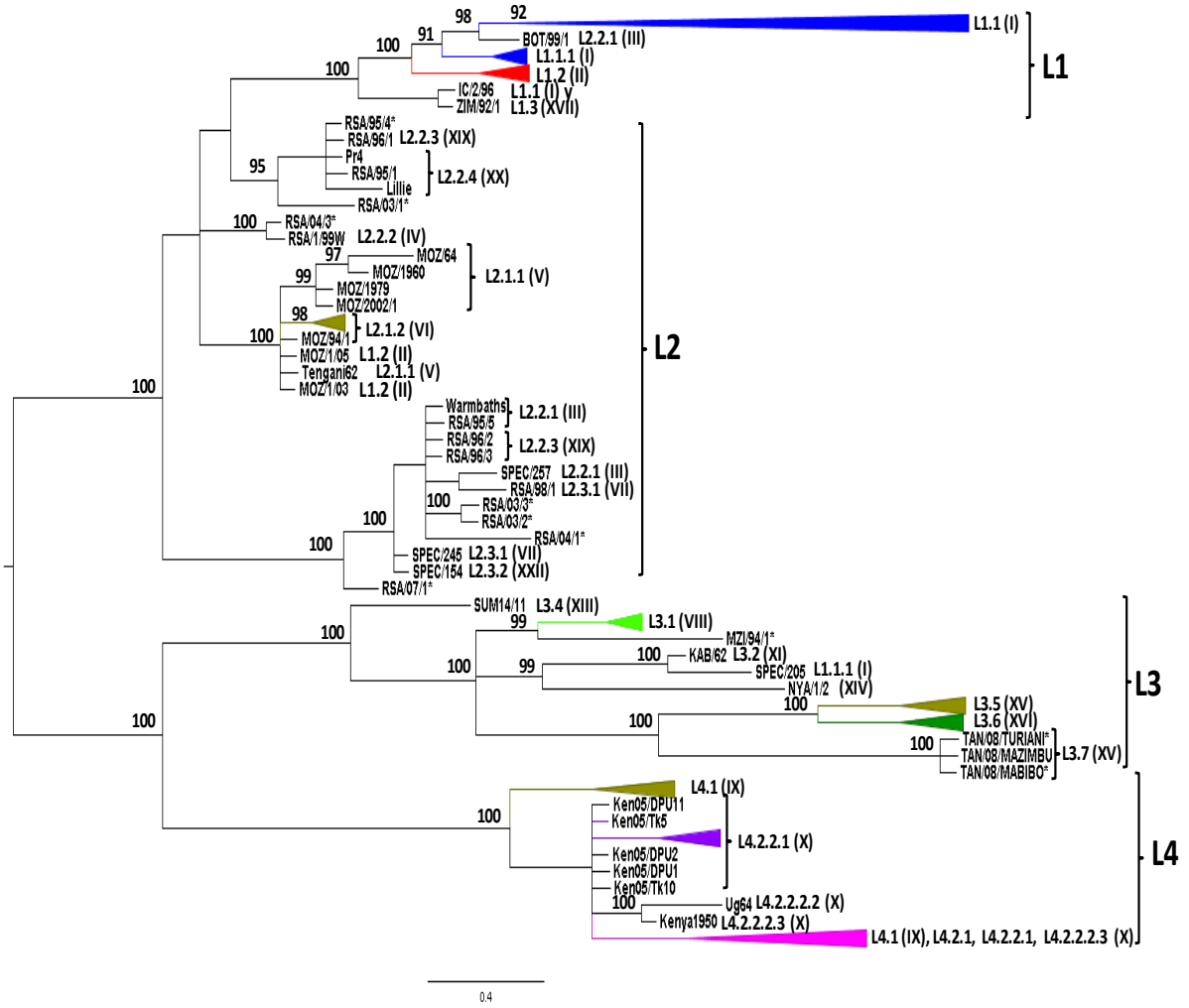


Figure 25 : Arbre phylogénétique en inférence bayésienne, construit selon le modèle HKY85 + Γ5 par le logiciel Mr Bayes avec les séquences du gène E183L. Les valeurs de bootstrap supérieures à 90 sont indiquées aux nœuds correspondants. L'arbre consensus a été généré après 2 X 10^6 générations, un échantillonnage chaque 1000 arbres et un burn-in des 50 X 10^5 premiers arbres construits (25%).

2-2-5- Reconstructions phylogénétiques utilisant le gène CP204L

2-2-5-1- Maximum de vraisemblance

Les critères d'information AIC, AICc et BIC ont proposés trois modèles évolutifs différents pour le jeu de données de séquences du gène CP204L : le modèle J1 + Γ5 + I, c'est-

à-dire un modèle prenant en compte les sites invariants des séquences ADN pour le critère AIC, le modèle HKY + Γ5 pour le critère AICc et enfin le modèle TN + Γ5 pour le critère BIC (Annexe 4). Les arbres ont été construits après 1000 ré-échantillonnages de l'alignement générés par le test LR-ELW. L'arbre sélectionné comme étant le plus en adéquation avec le jeu de données de séquences du gène CP204L a été celui construit avec le modèle HKY + Γ5 (Figure 26).

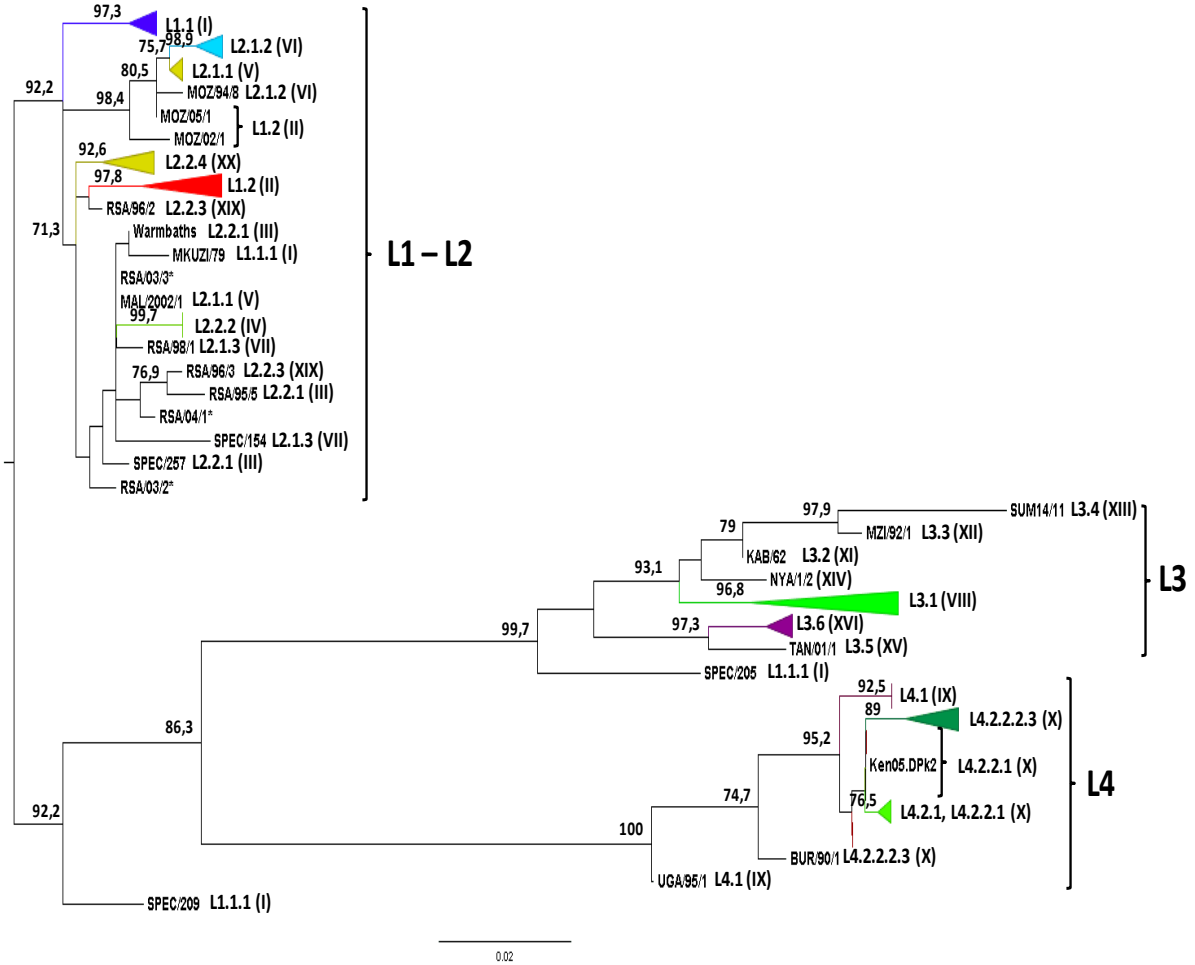


Figure 26 : Arbre phylogénétique décrivant les relations entre les séquences du gène CP204L des isolats du virus PPA. L'arbre a été construit en maximum de vraisemblance avec le modèle HKY85 + Γ5 proposé par le critère d'information AICc. Les isolats marqués d'un astérisque (*) ne participent pas à l'analyse du gène B646L qui a servi de support à la classification des virus PPA. Seules les valeurs de bootstrap supérieures à 70 sont indiquées.

Comme pour l'analyse des gènes B646L et E183L, les lignées L3 et L4 sont bien différenciées, tandis que les lignées L1 et L2 sont cette fois confondues au sein d'une même branche, confirmant la grande proximité biologique qui existe entre les isolats viraux qui les composent. L'isolat namibien SPEC/205 (L1-1-1) est, comme lors de l'analyse du gène E183L, replacé au sein de la lignée L3. L'isolat zambien NYA/1/2 est lui aussi également replacé au sein de la lignée L3. Un autre isolat namibien, SPEC/209 (L1-1-1), est isolé, même s'il se rapproche de la lignée L4. La lignée L1-1-1 semble donc davantage liée aux isolats d'Afrique australe qu'aux isolats d'Afrique de l'Ouest, d'Europe, des Caraïbes et d'Amérique du Sud, ce

qui fait sens au niveau géographique, même si l'isolat sud africain MKUZI/79 reste quant à lui toujours intégré à la lignée L1.

L'isolat MOZ/03/1 qui appartenait à la lignée L1-2 (génotype II, des isolats malgaches), se ségrége à l'intérieur de la lignée L2-1-1 (génotype V). De même que les isolats MOZ/05/1 et MOZ/02/1 se rapprochent des lignées L2-1-1 et L2-1-2 (génotypes V et VI). Cette discrimination est de fait géographiquement étayée, puisque l'ensemble des virus PPA formant ces deux lignées proviennent tous du Mozambique, et que le virus à l'origine de l'entrée sur l'île de Madagascar provient vraisemblablement de ce pays, situé sur le continent africain, face à l'île, alors que les échanges commerciaux entre ces deux pays sont établis.

La discrimination fine des isolats semble là encore difficile à établir au moyen d'une méthode phylogénétique par bifurcation. En effet, comme dans le cas des gènes B646L et E183L, des polytomies apparaissent lorsque les isolats sont très proches.

2-2-5-2- Inférence bayésienne

Tout comme avec les gènes B646L et E183L, deux modèles ont été appliqués pour la construction d'arbres en inférence bayésienne avec les séquences du gène CP204L, soit, le modèle HKY85 + $\Gamma 5$ sélectionné après l'analyse en maximum de vraisemblance, ainsi que le modèle GTR + $\Gamma 5$, le plus complexe, à titre comparatif. Dans les deux cas, l'analyse a été stoppée après atteinte d'un LRT inférieur à 0,01 au bout de 4,236 x 10^6 générations des MCMC avec le modèle HKY85 + $\Gamma 5$ et 4,583 x 10^6 générations avec le modèle GTR + $\Gamma 5$. Les lignées L1 et L2 sont confondues au sein d'une même branche, présentant de nombreuses multifurcations (Figure 27). Les lignées L3 et L4 sont bien différenciées, et soutenues par des valeurs de bootstrap égales à 100.

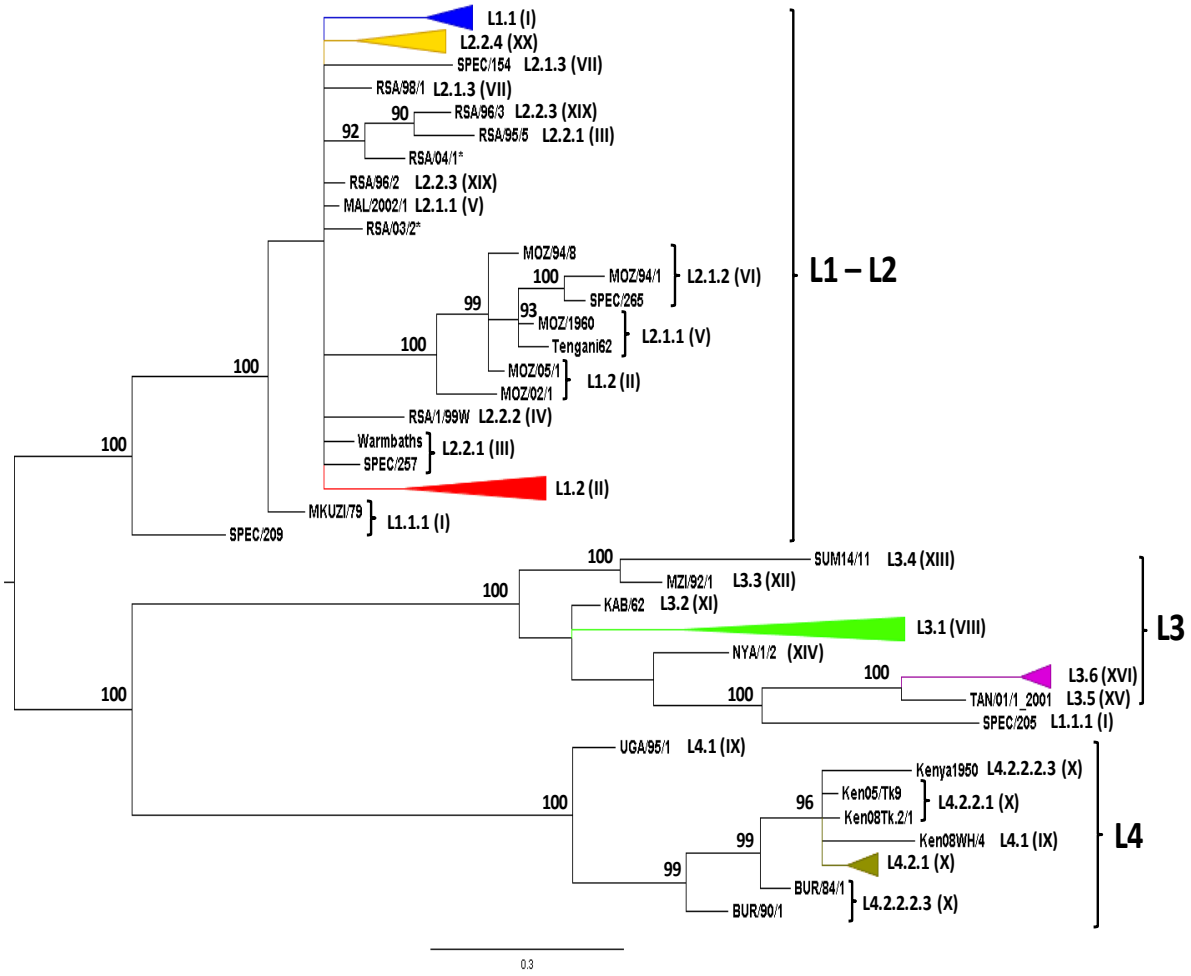


Figure 27 : Arbre phylogénétique décrivant les relations entre isolats de virus PPA construit en inférence bayésienne à partir des séquences du gène CP204L sous le modèle HKY85 + Γ5. Les isolats marqués d'un astérisque n'entrent pas dans l'analyse du gène B646L. Seules les valeurs de bootstrap supérieures à 90 sont indiquées.

3- Datation moléculaire

En datation moléculaire, il convient d'analyser l'évolution naturelle des séquences. Parmi les forces évolutives qui agissent sur les séquences, la pression de sélection positive induit des variations nucléotidiques plus rapides que l'évolution naturelle, c'est-à-dire le rythme des mutations aléatoires pouvant intervenir au fil du temps lors de la réplication des séquences d'ADN. Cette accélération évolutive va induire un biais dans la détermination du taux de substitutions μ, et donc un recul du TMRCA. Pour pallier ce biais, la pression de sélection s'appliquant sur les jeux de données de séquences que nous avons utilisés a été déterminée, et les codons sous pression de sélection positive ont été retirés des alignements de séquences avant de procéder à la datation moléculaire.

3-1- Détermination de la pression de sélection s'appliquant sur les séquences étudiées

L'analyse a été réalisée à l'aide de l'algorithme codeml du logiciel PAML version 4. Les arbres phylogénétiques choisis étaient ceux établis en maximum de vraisemblance avec les modèles HKY + $\Gamma 5$ pour les gènes B646L, E183L et CP204L. Ce sont donc ces arbres qui ont guidé la détermination du ratio $\omega = d_N/d_S$ des gènes étudiés. L'analyse a été conduite par site nucléotidique sur l'ensemble des séquences, sans tenir compte d'une possible variation du ratio selon les branches de l'arbre.

Concernant le gène B646L, le ratio ω a été déterminé à 0,223, ce qui signifie que le gène est soumis à une sélection faiblement positive. Sur l'ensemble du gène, 96.2% des codons montrent présentent un $\omega < 0{,}1584$. Parmi les 3.8% de codons restant, deux ont été détectés comme soumis à une pression de sélection positive : le codon 4, codant pour une histidine et le codon 28, codant pour une thréonine, avec $\omega = 1{,}540$ et $\omega = 1{,}407$, respectivement.

L'analyse du gène CP204L a déterminé un ω égal à 1,13, avec 98.8% des codons ayant un $\omega < 0{,}9$. Parmi les 1.2% des codons restant, seulement 3 ont été détectés avec un $\omega > 1$: il s'est agit du codon 31, codant le glutamate, du codon 123, codant pour une proline et du codon 176 codant pour une leucine. La valeur de ω pour ces trois codons est très élevée puisqu'elle a été déterminée à 6,80, 4,23, et 6,80, respectivement, ce qui explique la valeur relativement élevée de ω sur la totalité du gène.

Enfin, le calcul du d_N/d_S concernant le gène E183L a produit un $\omega = 0{,}286$ soit, comme pour le gène B646L, une faible pression de sélection positive. Cependant, si 74% des codons ont un $\omega < 0{,}236$, 26% d'entre eux présentent un ω élevé, dont 9 qui ont été détectés avec un $\omega > 1$: les codons 10 (tyrosine, $\omega = 1{,}46$), 23 (thréonine, $\omega = 1{,}44$), 100 (aspartate, $\omega = 1{,}26$), 104 (thréonine, $\omega = 1{,}46$), 122 (sérine, $\omega = 1{,}19$), 140 (proline, $\omega = 1{,}29$), 142 (valine, $\omega = 1{,}25$), 143 (glutamate, $\omega = 1{,}46$) et 149 (sérine, $\omega = 1{,}37$).

Ces codons ont ainsi été retirés des alignements de séquences des gènes correspondants pour ne pas biaiser le analyses de datation moléculaire. Les analyses suivantes ont donc été réalisées sur un alignement de 351 séquences de 393 nucléotides pour le gène B646L, 234 séquences de 453 nucléotides pour le gène E183L et 123 séquences de 534 nucléotides pour le gène CP204L.

3-2- Analyse en maximum de vraisemblance

L'analyse en maximum de vraisemblance a été réalisée à l'aide de l'algorithme du logiciel PAML version 4. Ce logiciel demande à la fois un alignement de séquences et un

arbre phylogénétique directeur construit en maximum de vraisemblance. L'ensemble du processus décisionnel pour le choix de l'arbre a donc été réitéré avec les alignements expurgés des codons soumis à pression de sélection.

Les arbres phylogénétiques choisis ont été ceux construits avec le modèle HKY + $\Gamma 5$ pour les gènes B646L, E183L et CP204L. L'ensemble des arbres a été construit après 1000 ré-échantillonnages effectué par le test LR-ELW. Pour les analyses contraintes par une horloge moléculaire (stricte ou locale), l'arbre a été enraciné comme décrit en 2-2-1 (Figure 12).

3-2-1- Test de l'hypothèse de l'horloge moléculaire stricte

L'arbre sans horloge moléculaire construit avec l'alignement du gène B646L a permis d'obtenir une valeur de vraisemblance LnB646L-1 = -1489,84. Avec l'arbre contraint par une horloge moléculaire stricte, on obtient LnB646L-2 = -1607,17, avec un taux de substitutions par site et par an $\mu = 0{,}000017$ +/- 0,000756 et un ancêtre commun datant de 405 pour une incertitude de 55449 ans.

Le test LRT du ratio des vraisemblances suit une distribution χ^2 avec n degrés de liberté. Sa valeur est égale au double de la différence entre les valeurs de vraisemblance de l'hypothèse nulle H_0 et de l'hypothèse testée H_1, soit $LRT = 2(lnH_1 - (\ln H_0))$. Dans ce test, l'hypothèse horloge moléculaire stricte représente l'hypothèse nulle H_0 et l'arbre construit sans contrainte l'hypothèse H_1. Le nombre de degrés de liberté correspond au nombre de paramètres indépendants entre les deux modèles. Sachant que nombre de séquences utilisées a été N = 351, qu'un arbre non enraciné (H_1) possède 2N-3 branches et qu'un arbre enraciné H_0 en possède N-1, le nombre de degrés de liberté que l'on doit prendre est N-2=349. Le résultat du LRT est donc égal à 234,66. Au seuil de 5%, la valeur du χ_c^2 est égale à 393,56, soit une valeur supérieure au LRT. Ainsi, l'hypothèse horloge moléculaire stricte ne peut être rejetée pour le gène B646L. Cependant, les incertitudes tant au niveau du taux de substitution que du TMRCA sont si élevées que le résultat du test LRT peut être remis en cause. L'horloge moléculaire stricte est de plus remise en cause car un test précédent, n'utilisant que les séquences uniques du jeu de données, avait rejeté sans aucun doute cette hypothèse.

La valeur de vraisemblance obtenue avec la construction de l'arbre E183L sans horloge moléculaire a été lnE183LH_1=-2365,02 tandis que la valeur du LR résultant de la construction de l'arbre contraint par l'horloge moléculaire stricte était lnE183LH_0=-2583,53. Le taux de substitutions par site et par an a été déterminé à 0,000040 +/- 0,000095. Le TMRCA a été fixé à 1054, avec une incertitude de 2241 ans. Le nombre de séquences utilisées pour construire ces arbres étant de 253 séquences, le nombre de degrés de liberté appliqués au test LRT a donc été de 251. La valeur du test LRT est donc de 437,02 ce qui est supérieur à la

valeur du χ_c^2 (288,95). L'hypothèse de l'horloge moléculaire stricte peut donc être rejetée au seuil de 5%.

Enfin, l'arbre construit sous l'hypothèse nulle H_0 avec le gène CP204L a produit une valeur de vraisemblance lnCP204LH_0=-2399,87 avec un taux de substitution 0,000004 +/- 0,000569, et un ancêtre commun datant de -12741 +/-1873229 ans. La valeur de LR générée avec l'arbre sous l'hypothèse H_1 a été lnCP204LH_1= -2323,60. Le résultat du test LRT a donc été de 152,54. Avec 121 degrés de liberté, la valeur du χ_c^2 est de 147,67. L'hypothèse horloge moléculaire est donc rejetée au seuil de 5%.

3-2-2- Horloge moléculaire locale

L'hypothèse de l'horloge moléculaire stricte a été rejetée pour les gènes E183L et CP204L, soit deux gènes sur les trois étudiés. Concernant le gène B646L, l'incertitude de mesure est telle que nous avons décidé de réaliser pour lui aussi une analyse utilisant une horloge moléculaire locale, c'est-à-dire autorisant la variation du taux de substitutions μ selon les branches de l'arbre. Pour ce faire, les branches « autorisées » à avoir μ indépendant doivent être indiquées dans la syntaxe de l'arbre.

Deux analyses ont été réalisées pour chaque gène, en autorisant différentes branches à avoir un μ indépendant. Les branches marquées ont en premier lieu été les branches majeures des arbres, représentant les quatre grandes lignées déterminées après la reconstruction phylogénétique par bifurcations et l'analyse en réseau. Pour l'analyse suivante, en plus de ces quatre grandes lignées, les lignées secondaires correspondant aux génotypes définis dans la littérature ont été autorisées à varier indépendamment. Pour le gène B646L, 4, puis 20 branches ont été ainsi autorisées à avoir un taux de mutation variable. Concernant les gènes CP204L et E183L, les arbres construits après avoir retiré les codons sous pression de sélection, ont montré seulement trois branches majeures, les lignées L1 et L2 étant regroupées au sein d'une même branche. Pour ces deux gènes, ce sont donc 3, puis, respectivement 10 et 20 branches qui ont été marquées.

Les résultats de la datation selon cette méthode ont été les suivants : les deux analyses concernant le gène B646L ont inféré un ancêtre commun approximativement aux mêmes dates, 812 puis 824, soit quatre siècles après celle inférée lors de l'analyse sous l'hypothèse de l'horloge moléculaire stricte. En revanche, des résultats générés pour les deux autres gènes ont été beaucoup plus disparates. Avec le gène E183L, la première analyse, dans laquelle trois branches étaient marquées a inféré un ancêtre commun en 700 tandis que la seconde l'a fixé en 1597 avant JC, alors que l'horloge moléculaire stricte le datait en 1054. Enfin, pour le gène CP204L, seule la première analyse a été menée à son terme, avec un Tmrca datant de 1407 avant JC. Lors de la seconde analyse, la valeur de vraisemblance

déterminée pour les arbres a atteint zéro avant la fin des calculs, et le Tmrca n'a donc pas pu être inféré par le logiciel.

Ces résultats montrent une grande hétérogénéité dans l'inférence des ancêtres commun, et ce quels que soient les gènes étudiés. L'influence des branches autorisées à avoir un taux de mutation variable est très importante sur l'inférence des Tmrca, engendrant une variation de plusieurs millénaires jusqu'à l'incapacité de la méthode à inférer l'âge d'un ancêtre commun. Déterminer de façon arbitraire quelles branches peuvent avoir un taux de variation variable, en se basant sur la topologie des arbres ou sur la signature génétique des isolats viraux, ne semble pas permettre une détermination optimale, ni même seulement précise, des Tmrca. La datation par la méthode du maximum de vraisemblance et de l'horloge moléculaire relâchée locale trouve sans doute là ses limites et ces résultats ne peuvent donc pas être considérés comme significatifs.

3-3- Analyse bayésienne par des chaines de Markov et technique Monte Carlo

3-3-1- Datation moléculaire du gène B646L

Les analyses bayésiennes ont été conduites sur 10^8 itérations, soit autant d'arbres générés. Les analyses, qui ont été effectuées après échantillonnage sur un total de 10^4 arbres, ont montré une convergence des probabilités postérieures pour chacun des paramètres testés, avec une distribution de ces probabilités autour de la moyenne du taux de substitutions et du TMRCA. De plus, aucune variation statistique significative des probabilités postérieures n'a été observée lorsque nous avons fait varier la valeur initiale de μ ou l'intervalle de distribution dans laquelle le taux de substitutions se situe.

Le taux de substitution a été déterminé à 6,49 x 10^{-4} substitutions par site et par an (Tableau 4), taux soutenu par une valeur d'ESS (« *effective sample size* ») égale à 2406,7. L'ancêtre commun le plus récent a été déterminé pour ce gène comme datant de 1712, avec un ESS de 323,11. Enfin, la valeur de vraisemblance a été calculée à LR=-1901,6 pour un ESS de 238,5. Pour valider un paramètre de l'analyse, l'ESS doit être supérieur à 200. Nos résultats peuvent donc être considérés comme statistiquement fiables.

L'arbre consensus construit après avoir rejeté les premiers 25% des arbres générés lors de l'analyse montre une ségrégation des isolats correspondant aux grandes lignées que nous avons définies précédemment (Figure 28). Les lignées secondaires sont peu ou prou les mêmes, même si la discrimination de la lignée L1-3 la place au sein de la lignée L1-1. La grande proximité moléculaire existant entre ces deux lignées secondaires explique ce rapprochement (Figure 20). Le TMRCA des lignées principales a pu être lui aussi déterminé. Ainsi, l'ancêtre commun des lignées L1 et L2 est daté de 1942, celui de la lignée L3 de 1965 tandis que celui de la lignée L4 a été estimé à 1892. La différence de TMRCA entre ces

différentes lignées, et principalement le recul de l'âge de l'ancêtre commun de la lignée L4 s'explique par le fait que les isolats appartenant à cette dernière proviennent de la région des grands lacs africains, berceau de la Peste porcine africaine. L'histoire évolutive des isolats viraux en provenance de cette région est donc logiquement plus longue.

La détermination de l'ancêtre commun le plus récent de deux autres lignées secondaires montre également un intérêt particulier : hormis l'isolat ZIM/92/1 (lignée L1-3) qui provient du Zimbabwe, les autres membres de la branche de l'arbre dans lequel il a été placé (lignée L1-1) proviennent d'Europe, d'Afrique de l'Ouest, des Caraïbes ou d'Amérique du Sud. L'ensemble de ces isolats est supposé émaner du premier virus PPA qui soit sorti d'Afrique pour atteindre le Portugal, à partir de l'Angola, en 1957. L'ancêtre commun le plus récent de cette lignée a été estimé à 1942, ce qui corrobore cette hypothèse. De même, il a été établi que le virus de la PPA était entré dans l'île de Madagascar en 1998 (Gonzague *et al.* 2001). La majeure partie des isolats formant la lignée L1-2 sont malgaches, issus des séquences que nous avons générées au cours de cette étude. Or, l'ancêtre commun le plus récent de cette lignée a été estimé à 1990. Là encore, la détermination des TMRCA est corroborée par l'histoire du virus dans les régions de prélèvements, étayant ainsi la véracité de nos analyses.

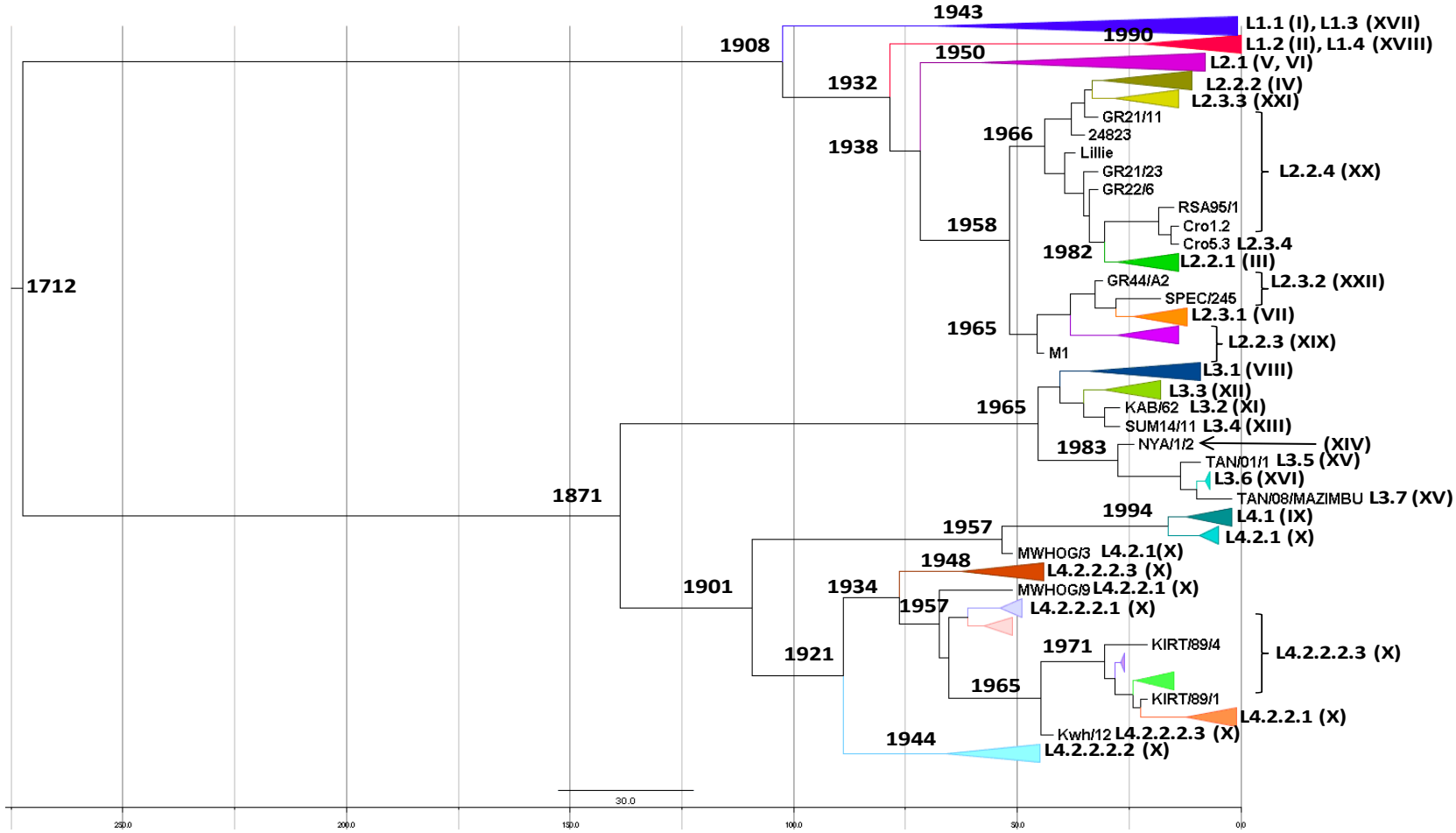


Figure 28 : Arbre phylogénétique daté des séquences du gène B646L construit par inférence bayésienne avec le logiciel Beast et le modèle HKY + Γ5. L'horloge moléculaire appliquée à l'analyse était une horloge relâchée lognormale non corrélée (UCLD). Les MCMCMC ont été tournées pendant 10^8 générations avec un échantillonnage chaque 10^4 arbre. L'arbre consensus a été généré après le rejet des 25 premiers pourcents des arbres générés, soit 2500 arbres.

3-3-2- Datation moléculaire du gène CP204L

Comme pour le gène B646L, les statistiques produites par l'analyse des 10^4 arbres échantillonnés sur les 10^8 arbres générés au total par les MCMCMC ont montré une bonne convergence des probabilités postérieures, sans qu'aucune variation significative n'ait été observée après modification de la valeur initiale et de l'intervalle de distribution du taux de substitution μ.

Le taux de substitution a été évalué à 6,64 x 10^{-4} substitutions par site et par an. L'ancêtre commun le plus récent des séquences du gène CP204L découlant de ce taux μ a été déterminé en 1700 (Tableau 4). Les ESS calculés pour ces deux valeurs ont été de 9248,15 et 569,22, soit très au dessus du seuil fixé à 200. La valeur de vraisemblance (-2397,9) possède elle aussi un support statistique très élevé de 3934,4. Pour cette analyse, tous les paramètres ont été validés sur le plan statistique.

L'arbre calculé à partir des 7500 arbres restant après le « burn-in » des 2500 premiers arbres générés ne montre pas la même organisation des quatre lignées principales d'isolats viraux déterminées avec le gène B646L (Figure 29). La lignée L4 est indépendante des trois autres, avec un TMRCA un peu plus récent, datant de 1921 (au lieu de 1892 lors de l'analyse du gène B646L). De même, les lignées L4-1 et L4-2 sont confondues au sein de cette branche de l'arbre. Les trois autres lignées sont discriminées à partir d'une seule branche. La lignée L3 reste elle aussi indépendante, mais replace en son sein l'isolat namibien SPEC/205, ségrégé dans la lignée L1-1 lors de l'analyse du gène B646L. Cette ségrégation est d'ailleurs géographiquement plausible, les isolats composant cette lignée provenant du Malawi, de la Zambie, de la Tanzanie, du Mozambique et du Zimbabwe. Cet isolat est de plus l'un des trois seuls isolats ne provenant par de l'Afrique de l'Ouest dans cette lignée L1-1.

Les lignées secondaires qui composaient les lignées L1 et L2 s'organisent quant à elles de façon différente : La lignée L1-1 est indépendante, ce qui, compte tenu de l'histoire des isolats qui la composent, supposés être la descendance de la souche virale ayant atteint l'Europe en 1957. L'ancêtre commun le plus récent de ces isolats est, de plus estimé, en 1953. La lignée L1-2 forme un clade indépendant, dont les seuls isolats ne provenant pas de Madagascar sont MUR/07/1, d'origine inconnue, et l'isolat géorgien Georgia2007. L'origine de l'épidémie survenue en Géorgie en 2007 pour ensuite se répandre dans le Caucase et jusqu'en Russie est de fait supposée être un virus en provenance de Madagascar. Le virus PPA est entré dans cette île en 1998, et l'ancêtre commun de ce clade est estimé, pour le gène CP204L, de 1992. L'isolat le plus proche du clade des virus PPA malgache est un isolat mozambicain (MOZ/02/2), le Mozambique étant vraisemblablement par les liens commerciaux unissant ces pays, à l'origine de la souche ayant atteint la grande île. Comme pour l'analyse du gène B646L, la détermination de TMRCA pour ces deux clades est soutenue par l'histoire épidémique du virus, confortant les l'ensemble des TMRCA estimés.

Les autres lignées secondaires composant les lignées L1 et L2 se ségrégent de façon moins précise que lors de l'analyse utilisant le gène B646L : les lignées L2-1-1 et L2-1-2 sont indiscernables, tout en formant un clade indépendant, tandis que les autres lignées secondaires déterminées forment un ensemble non réellement distinguable. Le nombre moins élevé d'isolats, seulement 123, utilisés lors de l'analyse du gène CP204L pourrait être à l'origine de cette discrimination plus évasive. Cependant, les TMRCA estimés pour les lignées L1-1 et L1-2, 1953 et 1992, respectivement, sont très proches de ceux estimés pour les mêmes clades lors de l'analyse du gène B646L (1942 et 1990), et même davantage fidèles aux données historiques. De plus, la proximité de l'ancêtre commun le plus récent pour ces deux gènes, à savoir 1712 et 1700, attestent de la fiabilité de nos analyses.

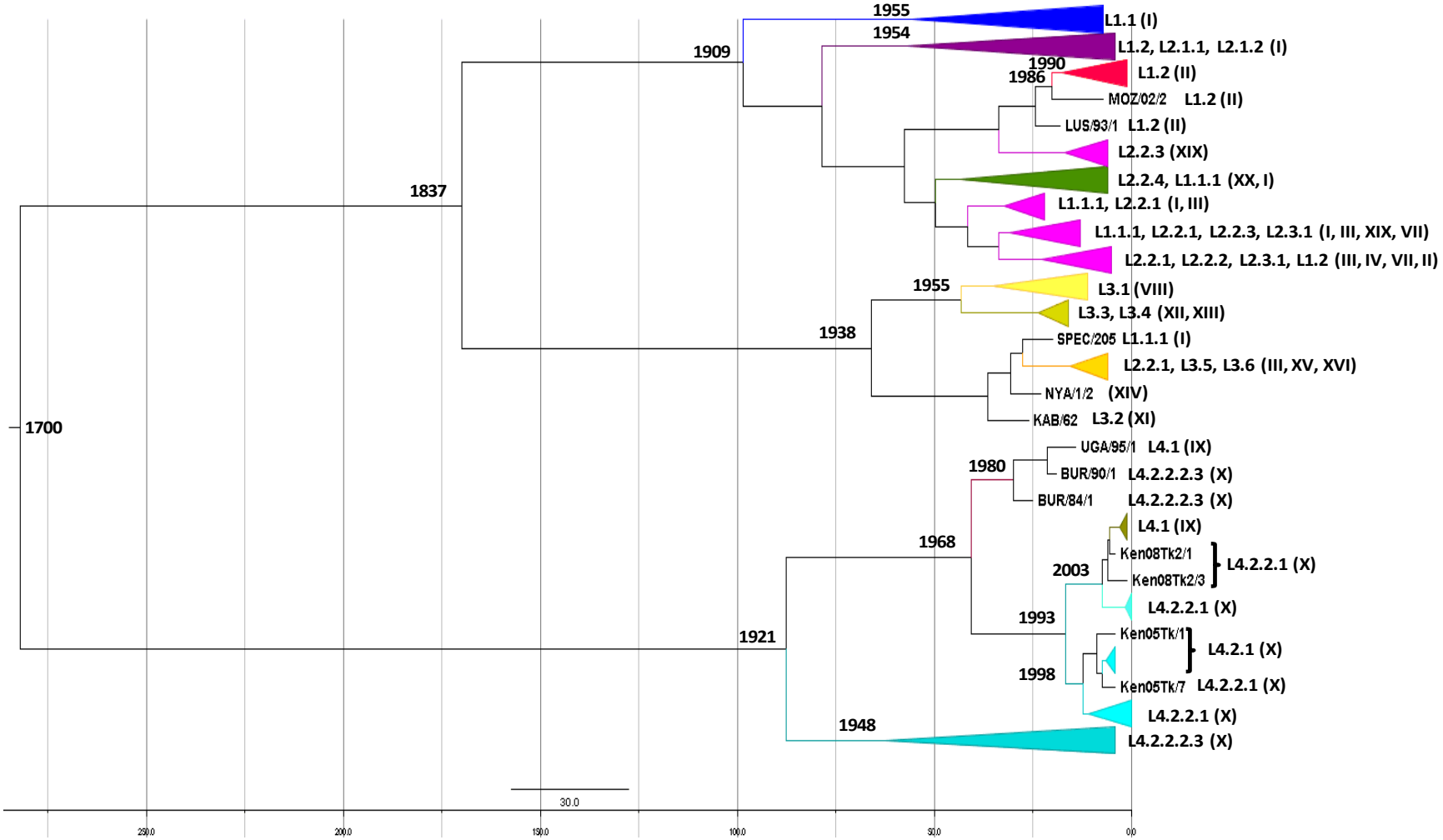


Figure 29 : Arbre généré par inférence bayésienne (modèle HKY + Γ5) avec le jeu de séquences du gène CP204L du virus PPA à l'aide du logiciel Beast. Les MCMCMC ont été tournées pendant 10^8 génération, avec échantillonnage chaque 10^4 arbre. L'arbre définitif a été généré après le rejet des 2500 premiers arbres, soit les 25 premiers pourcents.

3-3-3- Datation moléculaire du gène E183L

Le gène E183L, codant pour la protéine membranaire p54, a été celui montrant le plus de substitutions observées, ainsi que le plus grand nombre de codons soumis à pression de sélection positive, soit 9 codons. Après avoir retiré ces codons pour ne pas biaiser la détermination du taux de substitution μ, l'analyse en datation moléculaire effectuée par le logiciel BEAST a évalué μ à 2,7 x 10^{-4} substitutions par site et par an, résultat soutenu statistiquement par un ESS de 781,8. L'ancêtre commun le plus récent des séquences de ce gène a été déterminé en 1426, soit près de trois siècles avant ceux déterminés pour les gènes B646L et CP204L, pour un taux de substitution près de 2,5 fois inférieur (Tableau 4). Le résultat de ce TMRCA a été validé d'un point de vu statistique par un ESS égal à 914,3. La valeur de vraisemblance LR de l'analyse a été déterminée à -2161,9, pour un ESS de 1197. Comme pour l'analyse en datation moléculaire du gène CP204L, l'ensemble des paramètres de l'analyse ont été attestés par des ESS supérieurs à 200.

L'arbre daté consensus a été construit après le rejet des 2500 premiers arbres générés par l'analyse (Figure 30). Dans cet arbre, les lignées L3 et L4 sont, comme dans le cas du gène B646L, différenciées. Le TMRCA de la lignée L4 a été déterminé en 1891, soit à une date identique à celle obtenue lors de la datation des séquences du gène B646L. En revanche, l'ancêtre commun le plus récent de la lignée L3 à été fixé en 1885, soit 80 ans auparavant par rapport à l'analyse du gène B646L. La lignée L3 établie avec le gène E183L diffère de celle déterminée avec le gène B646L en plusieurs points. Tout d'abord, elle intègre l'isolat NYA/1/2 qui n'a pas pu être classé dans aucune des quatre lignées principales décrites pour le gène B646L, ni sur le plan de la signature moléculaire, ni par observation des branches de l'arbre phylogénétique déduit des données de séquences. Seules les analyses en réseau placent cet isolat sur le chemin menant à cette lignée (Figures 4 et 5). De plus, alors que la lignée L3-1 compte 45 représentants du gène B646L, sur un pas de temps de 18 ans (1983 – 2001), un seul représentant était disponible pour le gène E183L, l'isolat MwLil20/1, du Malawi, isolé en 1983. Il est à noter également que, si l'isolat NYA/1/2 est présent dans l'étude du gène CP204L, aucune séquence de la lignée L3-1 n'a pu être incluse dans l'analyse de ce gène. De même, il n'existe pas de représentant de la lignée L1-4 pour ce gène.

Comme pour l'analyse en datation moléculaire du gène CP204L, les isolats formant les lignées L1 et L2 avec le gène B646L ne forment plus deux branches distinctes. A l'intérieur de la lignée L1, certains isolats sont de fait replacés au sein de lignées secondaires différentes, comme l'isolat ivoirien IC/2/96 (lignée L1-1), qui se ségrége avec l'isolat ZIM/92/1 du Zimbabwe (lignée L1-3). Ces deux isolats se différencient de la lignée L1-1 d'Europe, d'Afrique de l'Ouest, des Caraïbes et d'Amérique de Sud, pour se rapprocher de la lignée L1-2 des isolats malgaches. Ce dernier est augmenté de deux isolats mozambicains (MOZ/02/1 et MOZ/02/2), d'un isolat zambien (LUS/93/1) ainsi que de l'isolat Georgia2007, déjà décrits comme appartenant à ce clade lors de l'analyse des gènes CP204L et B646L. Pour ce clade, l'ancêtre commun est daté de 1984, soit un recul dans le temps par rapport aux analyses de

datations moléculaire réalisées avec les deux autres gènes étudiés. L'ancêtre commun de la lignée L1-1 montre lui aussi un recul dans le temps, avec un TMRCA déterminé en 1893, soit un recul de 49 ans par rapport au gène B646L. Au final, les différents résultats obtenus sur les trois gènes sont rassemblés dans le (Tableau 4).

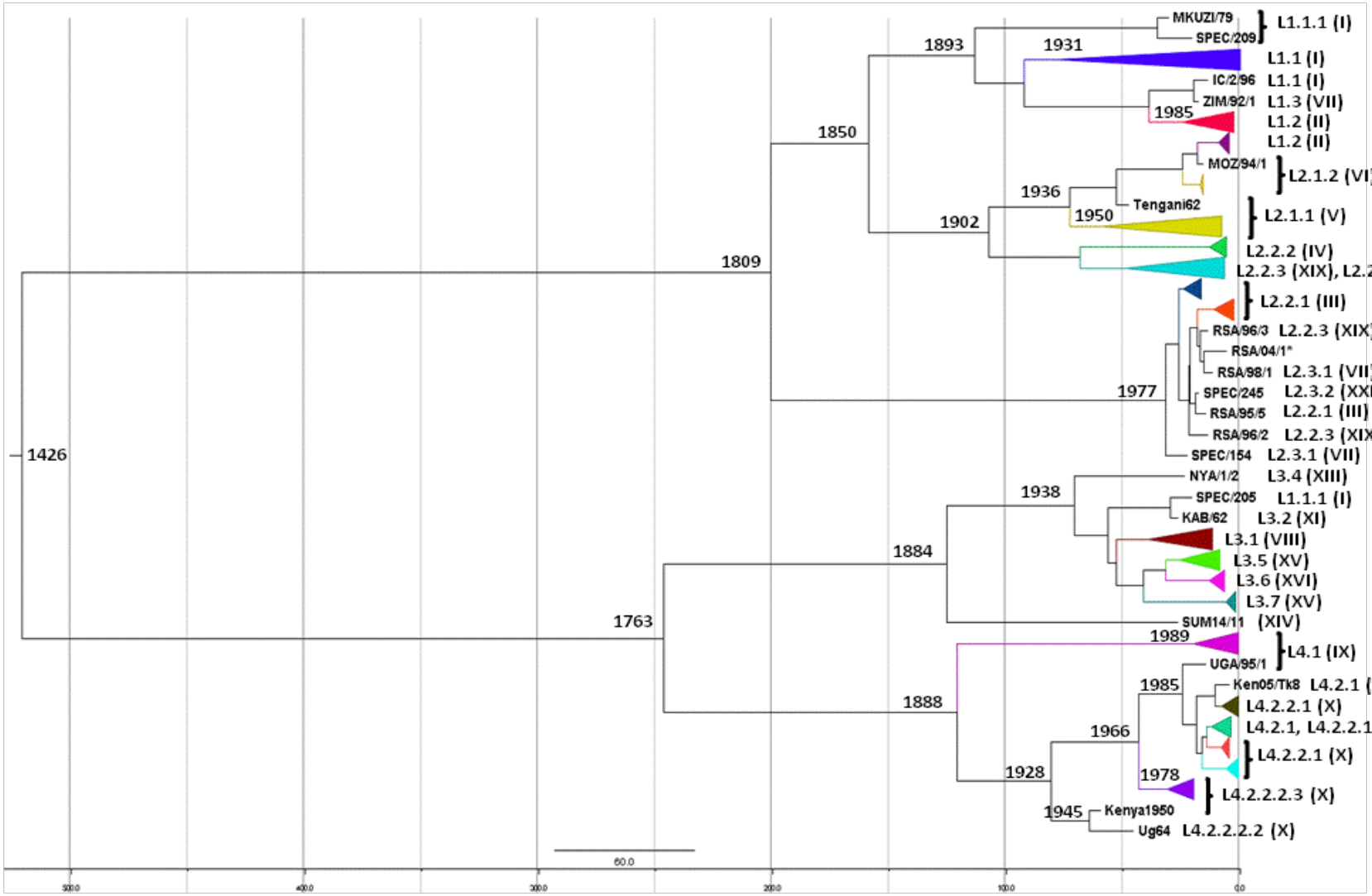


Figure 30 : Arbre phylogénétique daté du gène E183L obtenu par inférence bayésienne sous le modèle HKY + Γ5 à l'aide du logiciel Beast. Les MCMCMC ont été tournées pendant 10^8 générations, avec échantillonnage des arbres chaque 10^4 arbre généré. L'arbre consensus a été construit après le rejet des 2500 premiers arbres générés.

gène	horloge	Taille de la population	Modèle évolutif	LRT	Taux moyen de subst/site/an [95% HPD]	TMRCA [95% HPD]
B646L	strict	constant	HKY + Γ_5	-1993,8	$1{,}131 \times 10^{-4}$ [7,6 x 10-5 – $1{,}5 \times 10^{-4}$]	1622 [1466 – 1768]
	UCLN	constant	HKY + Γ_5	-1593,4	$6{,}9 \times 10^{-4}$ [$5{,}3 \times 10^{-4}$ – $9{,}13 \times 10^{-4}$]	1712 [1465–1894]
CP204L	strict	constant	HKY + Γ_5	-2453,07	$6{,}76 \times 10^{-5}$ [$1{,}9 \times 10^{-5}$ – $1{,}2 \times 10^{-4}$]	850 [-204 – 1610]
	UCLN	constant	HKY + Γ_5	-2398,03	$6{,}6 \times 10^{-4}$ [$5{,}57 \times 10^{-4}$ – $8{,}75 \times 10^{-4}$]	1700 [1422 – 1891]
E183L	strict	constant	HKY + Γ_5	-3902,7	$1{,}76 \times 10^{-4}$ [$9{,}1 \times 10^{-5}$ – $2{,}6 \times 10^{-4}$]	1529 [1243 – 1745]
	UCLN	constant	HKY + Γ_5	-2161,9	$2{,}7 \times 10^{-4}$ [$2{,}07 \times 10^{-4}$ – $4{,}03 \times 10^{-4}$]	1426 [1422 – 1720]

Tableau 4 : Récapitulatif des analyses et des résultats en datation moléculaire pour les trois gènes du virus PPA étudiés. 95% HPD (*highest posterior density*) : intervalle dans lequel sont comprises 95% des probabilités calculées.

DISCUSSION

Que l'objectif final soit d'établir une nomenclature permettant de classer des isolats viraux ou de déterminer l'horloge moléculaire qui prévaut à leur évolution, le choix des gènes d'intérêt s'avère fondamental en reconstruction phylogénétique. En effet, les mécanismes biologiques qui interviennent dans la diversité apparente entre des séquences d'ADN, ont un impact, tant sur la définition des liens existant entre les isolats, que sur la datation de l'ancêtre commun des souches virales. Or, la détermination du rythme évolutif propre aux séquences d'une lignée virale et de ses sous-lignées, ainsi que la caractérisation des relations unissant les virus entre eux, sont d'un grand intérêt, d'un point de vue épidémiologique et évolutionariste, mais aussi dans la définition d'une stratégie vaccinale adaptée. Ainsi, les forces évolutives à l'œuvre vont-elles modeler les séquences des pathogènes, en fonction des hôtes qu'ils infectent et de leur environnement. Il est désormais bien documenté que la nature des forces évolutives qui s'exercent sur les séquences géniques va impacter sur la détermination du taux de substitution qui les affecte. De même, la nature et/ou la fonction des protéines codées par les gènes d'intérêt (i.e. enzymes, protéines structurales, immunogènes…) induisent-elles la nature des forces qui s'appliquent sur elles. Ainsi, des protéines fortement immunogènes vont-elles être soumises à la pression du système immunitaire de l'hôte, induisant une sélection souvent diversifiante, quand les protéines structurales ou enzymatiques vont être soumises à une sélection négative (ou purificatrice) où la tendance sera plutôt à l'élimination des allèles défectueux (c'est-à-dire ne pouvant plus jouer leur rôle biologique), et ainsi diminuer la diversité entre les isolats. Par exemple, il a été montré qu'une sélection purificatrice forte pouvait résulter en une sous-estimation du TMRCA de lignées de virus à ARN, due à une mauvaise détermination de la longueur des branches entre les isolats (Wertheim & Pond 2011). De même, il est avéré par de nombreuses études que nombre de virus recombinent entre eux (Froissart *et al.* 2005 ; Heath *et al.* 2006 ; Kirkegaard & Baltimore 1986 ; Rhodes *et al.* 2003 ; Varsani *et al.* 2006), impliquant l'impossibilité de représenter leur évolution par un arbre unique et pouvant invalider le test de vraisemblance (LRT) calculé lors de la détermination de l'horloge moléculaire (Schierup & Hein 2000b). Les recombinaisons au sein des génomes viraux, qui plus est, ne se limitent pas à l'échange de gènes ou portions de gènes entre les virus eux-mêmes. De nombreux évènements de recombinaison apparaissent également entre le génome viral et le génome de son hôte. Ainsi, il est supposé que nombre de gènes immuno-modulateurs sont initialement des gènes de l'hôte que les virus se sont accaparés puisqu'ils leur confèrent un avantage évolutif. L'histoire évolutive de tels gènes est donc le fruit de plusieurs histoires parallèles ce qui sous-tend des relations très complexes entre les isolats.

Dans cette thèse, nous avons focalisé notre attention sur trois gènes du virus PPA. Le gène B646L, qui code pour la protéine majeure de la capside virale (VP72), et les gènes

CP204L et E183L, codant respectivement pour les protéines d'enveloppe p32 et p54. Tous les trois sont situés dans la partie conservée du génome du virus PPA, et les protéines pour lesquelles ils codent, induisent des anticorps chez le porc domestique (Neilan *et al.* 2004). Elles devraient donc être soumises à la pression du système immunitaire et ainsi à une sélection naturelle diversifiante. En effet, s'il n'inhibe pas la présentation des peptides par le complexe majeur d'histocompatibilité (CMH), le virus peut être détecté par le système immunitaire via les celllules cytotoxiques pré-existantes. Ces dernières vont alors induire les cellules infectées en apoptose, la mort cellulaire qui s'ensuit mettant fin au cycle de réplication du virus. Lors de la réponse cytotoxique, les antigènes viraux dégradés par protéolyse sont présentés aux cellules cytotoxiques associés aux protéines du CMH de classe I. Or, un récepteur de lymphocyte T cytotoxique ne reconnait qu'une seule association peptide viral/CMH. Par conséquent, si le virus a muté, l'association peptide viral/CMH ne sera plus reconnue, l'échappement du variant au système immunitaire laissant le temps nécessaire au virus de se répliquer et à l'infection de s'installer. Ainsi la sélection diversifiante favorisera-t-elle les variants, moins sujets à la réponse cytotoxique. Néanmoins, la localisation de la protéine VP72, dans le cœur même du virion, devrait la protéger d'une sélection purificatrice trop drastique. En effet, des modifications mêmes minimes de la protéine B646L peuvent s'avérer létale pour le virus, la capside virale étant fondamentale pour la morphogénèse du virus. Les contraintes, notamment sur la forme icosaédrique de la capside ne donnent que peu de latitude au virus pour muter.

Cette hypothèse est conforme aux observations que nous avons faites sur les codons placés sous pression de sélection positive, qui n'ont été qu'au nombre de deux pour le gène B646L, suggérant que la pression du système immunitaire n'a que peu d'impact en termes de diversification des séquences entre les isolats. Par conséquent, le taux de substitution par site et par an du gène de la VP72 devrait fournir des informations pertinentes sur l'évolution naturelle du virus. Le gène codant pour la protéine majeure de la capside virale a d'ailleurs déjà été utilisé dans des études phylogénétiques antérieures impliquant des virus de familles proches des *Asfarviridae* (Tidona *et al.* 1998), et également depuis une décennie pour déterminer les relations entre isolats de virus PPA. De même, la protéine p32, bien que localisée à l'enveloppe virale, est impliquée dans la traduction des gènes viraux via ses liaisons avec la protéine cellulaire hnRNP (Hernaez *et al.* 2008). Cette fonction est due à une structure tertiaire précise de la protéine, ce qui la rend susceptible aux mutations. De fait, seulement trois codons ont été détectés sous pression de sélection positive pour ce gène CP204L. Les gènes B646L et CP204L sont donc soumis à la sélection purificatrice, les mutations à l'intérieur de ces deux protéines pouvant s'avérer létales pour le virus. Il est à noter que la sélection purificatrice, en tout cas chez les virus à ARN, semble plus importante si le virus est transmis par un arthropode, sans que les causes en soient véritablement connues (Woelk & Holmes 2002). On peut supposer que le spectre d'hôtes impose au virus des contraintes sélectives élevées (Scott *et al.* 1994). A l'inverse, le système immunitaire semble avoir un réel impact sur l'évolution de la protéine p54 codée par le gène E183L, ce que montrent les 9 codons sous pression de sélection positive, ainsi que les évènements de

recombinaison détectés à l'intérieur des séquences. La p54 est une protéine traduite précocement et qui induit une forte réponse immune de l'hôte (Pastor *et al.* 1992). Au cours de l'adaptation des souches virales à la culture cellulaire, le poids moléculaire de cette protéine varie, le nombre de variants s'accroissant avec le nombre de passages en culture (Alcaraz *et al.* 1992). Ces changements sont observés aussi bien *in vitro* que *in vivo* (Pan & Hess 1985 ; Sumption *et al.* 1990), et se caractérisent majoritairement par un nombre variable de copies d'une séquence répétée de 12 nucléotides (Rodriguez *et al.* 1994), semblant constituer un mécanisme important de diversification du virus PPA. Nos résultats n'ont pas pris en compte ce mécanisme, puisque nous n'avons pas étudié la diversité induite par les indels, dont la variation de taille provient souvent de mécanismes de recombinaisons, mais seulement la diversité induite par les substitutions au sein de fragments de séquences de même longueur pour toutes les souches analysées.

Les recombinaisons, tout comme les codons sous pression de sélection positive, induisent une hausse du taux de substitution, un TMRCA faible, et une représentation phylogénétique des relations entre isolats semblable à celle résultant de l'analyse de séquences échantillonnées dans une population à croissance exponentielle (Schierup & Hein 2000a). Ayant retiré les séquences recombinantes ainsi que les codons sous pression de sélection positive, nous avons évité ce biais d'analyse. Un autre écueil dans l'analyse de l'évolution de séquences géniques est la possibilité de duplication des gènes au sein d'un génome. Ces gènes surnuméraires peuvent conduire à l'apparition de pseudo-gènes (souvent non fonctionnels) dont les histoires évolutives, indépendantes, entrainent un biais d'analyse. C'est la raison pour laquelle l'analyse du gène KP177R, codant pour la protéine d'enveloppe p22, un temps envisagé lui aussi en tant que gène d'intérêt, a été abandonnée après que la publication du génome complet de la souche Benin97/1 ait montré la présence de deux copies de ce gène, chacune portée par un des deux brins d'ADN du génome viral (Chapman *et al.* 2008) et pouvant induire le phénomène de paralogie cachée. La présence de duplicat des gènes B646L, CP204L et E183L n'a jamais été détectée au cours de l'analyse des 12 génomes complets de virus PPA séquencés à ce jour.

La phylogénèse du virus PPA s'est toutefois révélée complexe, au point qu'une représentation au moyen de méthodes strictement bifurcatives s'est avérée limitante pour expliquer les relations entre isolats. L'analyse en réseau des séquences géniques a d'ailleurs montré que les relations entre certains isolats ou groupes d'isolats ne suivaient pas le dogme voulant qu'un ancêtre ne produise que deux descendants, mais que des isolats pouvaient représenter des nœuds internes de l'arbre. De telles relations, où des isolats contemporains sont inférés en tant qu'ancêtres, et dans lesquelles il existe plusieurs chemins évolutifs permettant de mener d'un isolat à un autre ne peuvent être appréhendées finement par une phylogénie en bifurcation. Malgré ces relations complexes entre souches virales, les analyses en bifurcation comme celles en réseau ont toutes deux clairement montré que la clustérisation des isolats prenait la forme de quatre grandes lignées (L1 à L4), là où seulement trois avaient jusqu'alors été décrites (Boshoff *et al.* 2007). A l'intérieur de ces

lignées, la signature moléculaire des 22 génotypes décrits à ce jour a été établie, et deux nouvelles sous-lignées ont été déterminées. L'isolat Cro3.5, non caractérisé jusqu'ici, est la feuille unique d'une nouvelle branche de la lignée L2 (L2.3.4), tandis que l'isolat TAN/08/MAZILMBU, précédemment intégré au sein du génotype XV (lignée L3.5), s'en écarte pour se rapprocher du génotype VIII (lignée L3.1).

Si les signatures moléculaires des différentes lignées ont pu être établies, elles n'ont cependant pas toutes le même poids, car elles s'appuient sur un nombre inégal de substitutions nucléotidiques. Ainsi, la lignée L1 est-elle caractérisée au niveau de deux sites nucléotidiques spécifiques et quatre substitutions synonymes (G/T93 et A/G216), tandis que la lignée L4 est définie par douze sites et treize substitutions, dont trois ne sont pas synonymes, entrainant donc une mutation en termes d'acides aminés. Le génotype le plus représenté parmi toutes les séquences est la lignée L1.1 (génotype I), dont les isolats appartiennent pour la quasi-totalité au groupe dit ESAC – WA (Europe, South Africa, Caribbean and West Africa) qui fut introduit en Europe à partir d'Angola en 1957. Il est caractérisé par un seul nucléotide (A216), dont la substitution est synonyme en termes de protéine. La force de cette substitution, qui a été fixée dans les séquences malgré la circulation du virus sur trois continents et sur plus de soixante ans, pourrait s'expliquer en partie par le fait que le passage du codon GCG vers le codon GCA, même s'ils codent tous deux pour une alanine, montre une augmentation de la préférence codon déterminée pour le virus PPA (http://www.kazusa.or.jp/codon/) et s'accorde aussi au fait que le génome de ce virus est riche en AT. Ainsi, la substitution G → A se trouve-t-elle favorisée au cours de la réplication, permettant sa transmission à la descendance et donc sa fixation au sein de la lignée. La nouvelle nomenclature que nous avons proposée sur la base des signatures moléculaires, se substituant aux anciens génotypes et proposant des lignées supplémentaires, est également corroborée par les distances entre les clusters, qui permettent de fournir une double assise à cette classification.

La prise en compte des signatures moléculaires génétiques et des distances entre les géno-groupes dans le but d'établir une nomenclature a été utilisée pour classer d'autres virus comme celui de la maladie de Newcastle (Diel *et al.* 2012). Dans le travail de Diel *et al.*, il est précisé qu'il faut au moins quatre isolats présentant les mêmes caractéristiques pour créer un nouveau géno-groupe. Cette restriction à la caractérisation d'un nouveau génotype est soumise à un biais d'échantillonnage. En effet, moins de 4 souches portant des caractéristiques identiques ne signifie pas pour autant que ces souches ne constituent pas en soi une lignée à part entière. Nous avons été confrontés au même biais lors de l'établissement de notre classification des souches de virus PPA. L'échantillonnage des souches de virus PPA dans les régions où le virus suit un cycle selvatique a en effet permis d'isoler de nombreux virus dont la signature moléculaire est unique. Dans le cas de l'hypothèse d'une induction d'une diversité virale lors du passage du virus chez la tique molle, une souche unique peut constituer un phylum distinct qui peut ensuite évoluer au sein de la colonie d'arthropodes, sans que le virus ne se propage plus avant et ne soit à

l'origine d'un foyer. Il peut donc être abusif d'envisager qu'une nouvelle lignée ne puisse qu'être caractérisée en fonction du nombre d'isolats qui la représente, la divergence n'étant pas seulement affaire de nombre, mais aussi et surtout d'état de caractères.

Le virus PPA a montré un taux d'évolution plus élevé que chez d'autres virus à ADN (Duffy & Holmes 2008),. Ce haut taux d'évolution a produit un TMRCA récent, puisque l'émergence de l'ancêtre commun à toutes les souches contemporaines de virus PPA a été déterminée il y a environ trois siècles, autour de 1700. Il est communément admis que le berceau du virus PPA se situe en Afrique de l'est. En effet, la maladie a été décrite pour la première fois au Kenya en 1921, suivant un premier foyer épidémique en 1903. La maladie a ensuite diffusé de par le monde en suivant les axes commerciaux, pendant les décennies suivantes. Le grand pouvoir de diffusion ainsi que la virulence dont le virus fait preuve, indiquent que si l'émergence était advenue dans une autre région ayant des porcs domestiques, elle aurait été détectée. Originellement, le virus PPA est supposé être un virus de tiques (Plowright 1977) puisqu'il infecte les agarsides du genre *Ornithodoros*, à tous les stades de reproduction et avec un impact limité sur ces différents stades. Les tiques molles Ornithodoros sont des tiques endophiles. Cela signifie que leur habitat nécessite des conditions d'hygrométrie et de température stables. C'est la raison pour laquelle elles infestent les terriers où nichent les phacochères, formant des colonies. Etant, de plus, photophobiques, et prenant leur repas de sang rapidement (en comparaison avec les espèces de tiques dures), leur capacité à diffuser est très restreinte. Dans la nature, le virus entretient donc un cycle sauvage au cours duquel il est transmis horizontalement et verticalement au sein de la colonie de tiques (Hess *et al.* 1989 ; Plowright *et al.* 1974), et entre les tiques et les phacochères juvéniles qui vivent aux abords des terriers. Les juvéniles, qui naissent non infectés, montrent une virémie pendant deux à trois semaines suivant l'infection, pouvant ainsi compléter le cycle en réinfectant à leur tour des tiques lors du repas de sang (Plowright 1981 ; Thomson *et al.* 1980). Les adultes, quant à eux, ne présentent pas de virémie, même s'ils restent infectés à vie (Wilkinson 1989). Dans de telles conditions, le virus PPA n'est pas supposé diffuser largement autour des terriers de phacochères. La dérive génétique à l'œuvre chez le virus résulte donc en des ilots de diversité, représentés par ces terriers, où est maintenu un cycle selvatique ancien et lent, et qui connait peu d'entrées de nouvelles souches. Le cycle domestique que suit le virus, en revanche, est d'une autre nature. Dans ce cas, la contamination des porcs domestiques se fait essentiellement par contact ou ingestion de produits contaminés, rarement via la morsure d'une tique, et l'infection conduit dans la très grande majorité des cas à la mort très rapide de l'animal. Ce scenario est corroboré par les arbres phylogénétiques générés lors de cette étude, dans lesquels les lignées en provenance d'Afrique de l'est et du sud, régions où le virus suit un cycle selvatique, montrent davantage de diversité que les souches isolées lors de foyer épidémiques touchant des porcs domestiques, dans des régions où le cycle sauvage du virus n'existe pas. C'est également une des raisons qui explique que les virus de la lignée L1.1 aient peu dérivé alors qu'ils ont parcouru plusieurs continents sur plus de 60 ans. Si l'accès aux souches de virus est aisé en cas de foyers épidémiques, et donc au sein du cycle

domestique, la collecte de souches en provenance du cycle selvatique est compliquée par l'absence de symptômes et de mortalité chez les hôtes sauvages. La détection et l'isolement de nouvelles souches en provenance du cycle sauvage nécessitent donc un échantillonnage à grande échelle, dans des conditions de terrains qui rendent difficiles la conservation des échantillons. Dans de telles conditions, l'utilisation du papier filtre comme support de prélèvement, en offrant un accès facilité à de nouvelles souches, permettra de mieux résoudre les relations entre isolats, en affinant leur nomenclature.

La caractérisation de nouvelles souches est rendue difficile par l'absence ou le manque de séquences provenant des ascendants des souches circulantes. C'est le cas, par exemple, de la souche TAN/08/MAZIMBU, qui, collectée en Tanzanie durant une épidémie en 2008, avait été classée au sein du génotype XV (Misinzo *et al.* 2010), et qui, dans notre nomenclature, constitue la sous-lignée L3.7. En la plaçant au sein du génotype XV (L3.5), cette souche pouvait sembler, à tort, être la résurgence d'une souche plus ancienne, TAN/01/1, isolée durant un foyer épidémique tanzanien en 2001. La caractérisation des souches est également rendue difficile par le phénomène de recombinaisons entre différentes souches virales. La détection d'évènements de recombinaisons dans les séquences du gène E183L a conduit à retirer des analyses 16 isolats italiens. Malgré ces recombinaisons, ces isolats sont restés intégrés à la lignée L1, et plus particulièrement à la lignée L1.1 (génotype I), car possédant eux aussi une adénine en position 216. De plus, étant issus de la même région (Sardaigne), et montrant des zones de recombinaisons très semblables, la possibilité que ces souches aient émergées à partir d'un ancêtre commun recombiné ne doit pas être négligée. Dès lors, ils pourraient représenter une nouvelle sous-lignée de la lignée L1.1.

Pour consolider notre étude de la phylogénèse du virus PPA et la détermination d'un TMRCA fiable, trois gènes ont été analysés par deux méthodes différentes. Si l'analyse des trois jeux de séquences, en maximum de vraisemblance, au moyen du logiciel PAML, a permis de rejeter l'hypothèse d'horloge moléculaire stricte comme prévalant à l'évolution du virus, elle n'a, en revanche, pas permis d'obtenir des résultats cohérents en ce qui concerne l'estimation de l'âge de l'ancêtre commun lorsqu'une horloge relâchée locale a été utilisée. A l'inverse, l'approche bayésienne a généré des résultats congruents. Avec cette méthode, l'analyse des gènes B646L et CP204L a déterminé un ancêtre commun autour de 1700 tandis que l'analyse du gène E183L le situait trois siècles auparavant, autour de 1400, et estimait un taux de substitutions par site et par an deux fois inférieur aux deux autres gènes. Ce TMRCA reculé dans le temps peut tout à fait s'expliquer par ce taux de substitution inférieur conjugué au taux élevé de substitutions non synonymes et synonymes que nous avons observé dans le jeu de données de séquences correspondant. Comme l'a montré l'analyse du d_N/d_S, le système immunitaire influence l'évolution de la séquence de ce gène. Ainsi, l'impact de la pression de sélection sur le gène E183L se traduit par davantage de variabilité, ce qui induit un TMRCA plus ancien. Le rythme évolutif naturel du virus de la PPA devrait donc être mieux représenté par les gènes B646L et CP204L. De fait, deux

éléments permettent d'avoir confiance en la date du TMRCA calculée pour ces deux derniers gènes, autour de 1700. En effet, l'analyse des gènes B646L et CP204L a déterminé des TMRCA en 1943 et 1955, respectivement, pour la lignée L1.1 et en 1990 pour la lignée L1.2 dans les deux cas. Or, la lignée L1.1, quasi intégralement composée d'isolats provenant du groupe ESAC – WA, est supposée avoir émergée à la fin des années 1950 (Bastos *et al.* 2003). La lignée L1.2 inclut en majorité des isolats provenant de Madagascar, la PPA ayant été introduite pour la première fois dans l'île en 1998 (Gonzague *et al.* 2001). Les taux de substitutions par site et par an que nous avons déterminés au cours de cette étude ont été plus élevés qu'attendus, notamment par comparaison avec d'autres grands virus à ADN double brins tel que le gamma-herpes virus de vertébrés (10^{-9} subst./site/an), ou même des petits virus à ADN double brins comme le polyomavirus JCV (10^{-7} subst./site/an) (Duffy & Holmes 2008). Ainsi, avec un taux d'évolution compris entre 10^{-4} et 10^{-5} subst./site/an, le virus PPA se rapproche du taux d'évolution des virus à ARN, qui se situe, quant à lui entre 10^{-2} et 10^{-5} subst./site/an (Hanada *et al.* 2004).

Le virus PPA, comme les autres grands virus à ADN double brins, et d'autant plus qu'il est asymptomatique chez les suidés sauvages africains, est supposé avoir co-évolué avec ses hôtes (Holmes 2004). Une coévolution signifie une longue et ancienne histoire du virus dans la nature, c'est-à-dire, dans le cas qui nous occupe, un cycle selvatique établi depuis longtemps. Cependant, un taux d'évolution élevé couplé à un ancêtre commun récent ne semble pas compatible avec une ancienne coévolution entre les virus et ses hôtes. En effet, une telle association devrait logiquement induire un taux d'évolution bas et un TMRCA reculé dans le temps (Holmes & Drummond 2007). Or, il s'avère également que, pour un virus qui se réplique à haut niveau à l'intérieur de ces hôtes, un taux faible de substitutions par site et par réplication peut tout de même générer une accumulation de diversité qui conduit à déterminer un taux élevé de substitutions par site et par an (Hughes *et al.* 2009). Cette caractéristique a été décrite chez les virus virulents induisant une forme aigüe de maladie et qui possèdent alors un taux élevé d'évolution en nombre de substitutions par site et par an (Firth *et al.* 2010). A l'inverse, une affection asymptomatique de l'hôte avec une réplication à bas niveau devrait conduire à un taux de substitution par site et par an lui aussi diminué. Le virus PPA possède ces deux caractéristiques, à savoir asymptomatique chez les suidés sauvages africains tandis qu'il est hautement contagieux et létal pour le porc domestique. On pourrait voir une contradiction entre le taux de réplication du virus chez le porc domestique et l'apparente conservation des virus dans la lignée L1.1, comparativement à la grande diversité observée parmi les virus circulant dans les régions où existe un cycle selvatique. Cependant, la réplication asymptomatique du virus chez les suidés sauvages d'Afrique, au moins chez les juvéniles, est relativement importante, avec des titres atteignant 10^{2} dans le sang circulant et $10^{6,6}$ dans les ganglions lymphatiques (Wilkinson 1989). Associée à une possible diversité générée chez la tique dans laquelle au stade adulte, le virus se réplique à un titre de 10^{4} à 10^{6} HAD_{50}/mg selon les organes (Kleiboeker & Scoles 2001), et un maximum observé de $10^{4,3}HAD_{50}$/tique (Basto *et al.* 2006), on aurait alors dans

le cycle selvatique les conditions idéales pour qu'un virus à ADN double brin évolue relativement vite.

Un TMRCA datant de 300 ans, doit donc être la conséquence d'un évènement inattendu qui serait survenu dans le berceau de l'infection selvatique et qui aurait permis l'émergence de l'ancêtre commun aux virus PPA contemporains. Notre hypothèse est que cet évènement fait suite à l'introduction du porc domestique en Afrique. Le porc domestique n'est pas originaire d'Afrique sub-saharienne. En effet, il tient son ascendance du sanglier sauvage d'Eurasie et d'Afrique du nord, à partir duquel il a été sélectionné au cours du temps (Gifford-Gonzalez 2011). Bien que des traces archéologiques de l'introduction de porcs domestiques entre le 3ème et le 7ème siècle aient été découvertes en Afrique du Sud (Plug 2001), le porc domestique ne devait pas faire partie des cheptels sud et est africains à cause du mode de vie des bergers de cette époque (Swart 2010). Ces derniers étaient des nomades, or, le porc domestique n'est pas adapté au nomadisme, encore moins aux transhumances. Le porc domestique a été introduit pour la première fois en Afrique par les chinois il y a quelques 600 ans (Levathes 1994), puis à nouveau par les portugais il y a entre 300 et 400 ans (Blench 1999), durant leur période expansionniste, dans le but d'explorer et conquérir de nouveaux territoires leur permettant d'accroitre leur capacité de négoce. L'hypothèse d'une introduction à partir d'Europe et d'extrême Orient a d'ailleurs été confirmée par des études phylogénétiques qui ont mis au jour les deux origines dans le patrimoine génétique des porcs locaux africains (Ramirez *et al.* 2009). Faisant suite à la circumnavigation du continent africain par les européens au cours des XVème et XVIème siècles, les colonisateurs ont introduit l'élevage du porc domestique durant les XVIème et XVIIème siècles (Swart 2010). Principalement apportés sur les côtes est africaines par les portugais en provenance de Goa, les porcs domestiques ont alors diffusé, à partir du Mozambique vers le nord en direction de la corne de l'Afrique (Blench 1999). Si les portugais ont conquis le Kenya, ce n'était pas pour y installer des colonies de peuplement, mais pour développer des comptoirs commerciaux sur la route des Indes, où ils s'étaient déjà établis depuis la fin du XVème siècle, à la suite des voyages exploratoires de Vasco de Gamma. Ils restèrent au Kenya jusqu'à leur défaite face aux Arabes en 1698, puis quittèrent définitivement le pays en 1720. En dépit de la présence des Arabes et du tabou alimentaire représenté par le porc pour ces peuples, la consommation de viande porcine n'a jamais cessé au sein de certains groupes ethniques, tels que les Waata, installés au Kenya depuis le XVIème siècle, et appelés par les autres ethnies les « Walyankuru », c'est-à-dire « ceux qui mange du porc » (Kusimba 2000). La continuation de l'utilisation, même limitée, du porc à des fins alimentaires aura permis au virus de continuer lentement et silencieusement de diffuser parmi les espèces de porcs sensibles à la maladie. A partir de 1873, le Kenya fut progressivement annexé par les britanniques jusqu'en 1885, britanniques qui, suite à une épidémie de peste bovine à la fin du XIXème siècle, introduisirent massivement le porc domestique pour en faire l'élevage, d'abord en provenance des Seychelles en 1904, puis du Royaume Unis en 1905. A cette époque, les porcs étaient principalement élevés en divagation, en contact avec la faune sauvage locale premier foyer épidémique en 1907. A

partir de là, le virus aurait tout loisir de diffuser à la faveur des échanges commerciaux, compte-tenu de sa capacité de persistance dans l'environnement et ses multiples modes de transmission.

DISCUSSION GÉNÉRALE CONCLUSION ET PERSPECTIVES

Dans le cadre des maladies ayant un fort pouvoir de diffusion, comme c'est le cas de la Peste porcine africaine, la circonscription de la maladie repose avant tout sur un diagnostic rapide. Cependant, le diagnostic ne survient qu'après la déclaration d'un foyer infectieux. Or, il convient également de caractériser les souches virales afin de pouvoir les tracer, surveiller leur circulation et adapter les mesures de contrôle (Zollner 2004). En effet, détection, caractérisation et épidémio-surveillance sont les bases d'un contrôle efficace de la maladie. Pour être fiable, ce triptyque doit reposer sur l'analyse de nombreux échantillons, dont l'acheminement et le stockage sont très contraints. D'une part, les conditions environnementales où sévit la PPA, à savoir très majoritairement des pays tropicaux en développement, sont assez extrêmes pour la conservation des échantillons biologiques et la chaine du froid est bien souvent difficile à maintenir sur des déplacements assez longs pour accéder aux animaux. D'autre part, les restrictions internationales sur le transport des matières dangereuses, incluant notamment les produits infectieux et la carboglace, renchérissent le coût du transport et freinent l'échantillonnage de matériels. Par ailleurs, lorsque le virus PPA a été introduit dans région indemne, il est très difficile de l'éradiquer, ne serait-ce qu'à cause de la présence de réservoirs sauvages (tiques molles et porcs sauvages) ou de porcs domestiques en divagation, l'absence de vaccin ou encore la faiblesse des structures vétérinaires de surveillance et de traitement des prélèvements.

C'est dans ce contexte et l'émergence de la PPA à Madagascar en 1998 puis dans le Caucase en 2007, que nos travaux de recherche ont été entamés dès 2007.

Le premier objectif de cette thèse, à savoir le développement d'un outil permettant le transport et la conservation de prélèvements sanguins dans des conditions climatiques difficile, a été réalisé par la mise au point d'une PCR directe à partir de papier filtre imbibé de sang circulant. Cette méthode a déjà été employée pour détecter de nombreux virus humains tels que le virus de l'immunodéficience humaine (VIH) (Beck *et al.* 2001 ; Yourno & Conroy 1992), le virus de la rougeole (Katz *et al.* 2002 ; Mosquera Mdel *et al.* 2004) ou encore celui de l'hépatite C (Abe & Konomi 1998), ainsi qu'un panel de virus animaux à ARN ou ADN génomique (Dubay *et al.* 2006 ; Wang *et al.* 2002). Cependant, elle requiert normalement l'extraction du matériel nucléique avant de procéder à l'amplification génique puis au séquençage. Ici, nous avons développé un test où le papier filtre imbibé de sang et conservé à température relativement élevée, peut être utilisé directement, sans traitement préalable. En cela, notre objectif consistait à réduire les coûts de diagnostic pour les laboratoires du sud. Nos résultats valorisés sous forme de publication, sont désormais exploités sur le terrain pour effectuer une surveillance épidémiologique dans plusieurs pays

d'Afrique (publication en préparation à laquelle nous sommes associés). En plus d'être facile d'emploi, le support buvard pour le prélèvement de sang est très peu onéreux et donc particulièrement bien adapté aux pays en développement. S'il ne règle pas la question de l'inactivation du virus comme le revendiquent d'autres matériaux comme les cartes FTA (GE Healhcare/Whatman), ils permettent cependant de s'affranchir de l'utilisation de la carboglace, considérée comme substance dangereuse pour le fret aérien, dans le cadre du transport vers les laboratoires internationaux de référence pour la confirmation de la maladie. En outre, l'absence d'inactivation du virus et la démonstration récente qu'il était possible de réisoler le virus à partir du papier buvard que nous utilisons (publication en préparation à laquelle nous sommes associés), rendent ce support attractif pour envisager, au-delà d'un génotypage rapide direct sur buvards, l'amplification du matériel biologique pour une caractérisation plus fine du virus.

La caractérisation des souches associées aux foyers est un moyen très important de tracer les sources possibles de la contamination et ainsi aider à la prise des mesures sanitaires les plus adéquates. La connaissance de la souche à l'origine de foyers, comme se fut le cas dans le Caucase, a en effet permis de déterminer la route empruntée par le virus pour atteindre la Géorgie. Notre hypothèse forte est la voie du commerce maritime entre Madagascar et l'Europe et l'utilisation des eaux grasses des navires pour alimenter des porcs domestiques situés à proximité des ports. Ainsi des mesures sanitaires spécifiques peuvent-elles être mises en place pour lutter contre ces portes d'entrées de la maladie. L'intérêt d'une caractérisation fine des isolats prend donc ici tout son sens. Les virus PPA ont commencé à être caractérisés au début des années 2000 (Bastos *et al.* 2003), et depuis, l'addition dans les analyses phylogénétiques de nouveaux gènes et de nouvelles souches n'a cessé de faire évoluer la nomenclature dans le sens d'une plus grande diversité (Boshoff *et al.* 2007 ; Lubisi *et al.* 2005, 2007). Néanmoins, la génération de cette diversité n'avait jusqu'ici jamais été étudiée sur le plan évolutif. Comprendre l'évolution du virus au niveau moléculaire procure un double avantage. D'abord la façon dont le virus se diversifie a un impact sur le choix de la stratégie vaccinale la mieux adaptée au virus. Ensuite la détermination du rythme évolutif des souches permet de mieux comprendre leur filiation, c'est-à-dire leur(s) origine(s), et donc leur diffusion.

Nous avons déterminé au cours de cette thèse que le virus PPA évoluait avec un rythme supérieur aux autres virus à ADN génomique double brins (Drake *et al.* 1998 ; Drake & Hwang 2005), et dans la limite basse des virus à ARN (Drake 1993). Ce taux de substitution élevé est associé à la détermination d'un ancêtre commun proche dans le temps, puisque datant du début du XVIIIème siècle, bien que le virus soit supposé avoir co-évolué sur une longue période avec ces hôtes vertébrés (suidés sauvages africains) et invertébrés (tiques molles du genre *Ornithodoros*). Cette émergence récente s'accorde cependant bien avec l'histoire de l'introduction en Afrique de l'Est du porc domestique européen, si l'on considère l'hypothèse que l'ancêtre commun aux virus PPA contemporains préexistait dans cette région, maintenu de façon discrète dans un cycle de réplication selvatique impliquant

tiques molles et suidés sauvages d'Afrique. Le virus PPA appartient à la famille des grands virus à réplication nucléocytoplasmique (NCLDV) dans laquelle il représente la famille des *Asfarviridae* (Iyer *et al.* 2006). Cependant, si plusieurs études ont analysés les relations qui unissent les 7 familles de NCLDV (Iyer *et al.* 2001 ; Koonin & Yutin 2010), aucune ne s'est intéressée à leur taux d'évolution. Le virus PPA étant membre unique des *Asfarviridae*, il serait intéressant de comparer son rythme évolutif avec celui des virus de parentés proches afin de déterminer si, au-delà des caractéristiques phénotypiques et réplicatives qu'ils partagent, les virus NCLDV montrent un rythme évolutif différent des autres virus à ADN double brins, et quel est le positionnement du virus PPA dans cette dynamique. La phylogénèse du virus PPA a été effectuée à partir de séquences d'acides nucléiques disponibles relativement courtes et à partir de souches majoritairement isolées sur des porcs domestiques lors de foyers épidémiques. Or, une précision supérieure des analyses phylogénétiques, de la détermination du taux de substitutions et de la datation moléculaire peut être atteinte avec un échantillonnage plus conséquent (McCormack *et al.* 2009) et de plus longues séquences nucléiques (Bromham *et al.* 2000). Aussi, nos résultats pourront-ils être affinés par l'utilisation de séquences de génomes complets de virus, auxquelles les nouvelles stratégies de séquençage comme les NGS (*Next Generation Sequencing*) donnent aujourd'hui un accès plus aisé, permettant d'inclure un nombre sans cesse croissant d'isolats viraux. En plus d'appartenir aux NCLDV, le virus PPA est le seul arbovirus à ADN génomique détecté à ce jour. L'hypothèse d'une diversité induite par les tiques (Dixon & Wilkinson 1988) n'a cependant jusqu'ici jamais été vérifiée. Il conviendra donc, pour mieux décrypter l'histoire et la dynamique évolutive du virus PPA d'établir un échantillonnage efficace de souches à partir des hôtes sauvages du virus, et d'étudier la diversité en analysant les génomes complets des virus ainsi isolés. Il est également envisagé comme suite à ce travail d'étudier la génération de cette diversité lors du passage du virus en conditions laboratoires dans le vecteur tique.

BIBLIOGRAPHIE

Abe, K., Konomi, N. (1998). Hepatitis C virus RNA in dried serum spotted onto filter paper is stable at room temperature. *J. Clin. Microbiol*. **36**, 3070-3072.

Afonso, C. L., Alcaraz, C., Brun, A., Sussman, M. D., Onisk, D. V., Escribano, J. M., and Rock, D. L. (1992). Characterization of p30, a highly antigenic membrane and secreted protein of African swine fever virus. *Virology* **189**(1), 368-73.

Afonso, C. L., Piccone, M. E., Zafutto, K. M., Neilan, J., Kutish, G. F., Lu, Z., Balinsky, C. A., Gibb, T. R., Bean, T. J., Zsak, L., and Rock, D. L. (2004). African swine fever virus multigen family 360 and 530 genes affect host interferon response. *J. Virol.* **78**(4), 1858-1864.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716-723.

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S., Fukushima, T. (1960). On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Jpn J. Microbiol.* **4**, 219-227.

Alcami, A., Angulo, A., Lopez-Otin, C., Munoz, M., Freije, J. M., Carrascosa, A. L., and Vinuela, E. (1992). Amino acid sequence and structural properties of protein p12, an African swine fever virus attachment protein. *J. Virol.* **66**(6), 3860-8.

Alcaraz, C., Brun, A., Ruiz-Gonzalvo, F., Escribano, J.M. (1992). Cell culture propagation modifies the African swine fever virus replication phenotype in macrophages and generates viral subpopulations differing in protein p54. *Virus Res.* **23**, 173-182.

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *StaT. Sci.* **12**, 162-176.

Almazan, F., Rodriguez, J. M., Andres, G., Perez, R., Vinuela, E., and Rodriguez, J. F. (1992). Transcriptional analysis of multigene family 110 of African swine fever virus. *J. Virol.* **66**(11), 6655-67.

Almazan, F., Rodriguez, J. M., Angulo, A., Vinuela, E., and Rodriguez, J. F. (1993). Transcriptional mapping of a late gene coding for the p12 attachment protein of African swine fever virus. *J. Virol.* **67**(1), 553-6.

Alonso, F., Dominguez, J., Vinuela, E., and Revilla, Y. (1997). African swine fever virus-specific cytotoxic T lymphocytes recognize the 32 kDa immediate early protein (vp32). *Virus Res.* **49**(2), 123-30.

Anderson, E. C., Hutchings, G. H., Mukarati, N., and Wilkinson, P. J. (1998). African swine fever virus infection of the bushpig (Potamochoerus porcus) and its significance in the epidemiology of the disease. *Vet. Microbiol.* **62**(1), 1-15.

Andres, G., Alejo, A., Salas, J., and Salas, M. L. (2002a). African swine fever virus polyproteins pp220 and pp62 assemble into the core shell. *J. Virol.* **76**(24), 12473-82.

Andres, G., Garcia-Escudero, R., Salas, M. L., and Rodriguez, J. M. (2002b). Repression of African swine fever virus polyprotein pp220-encoding gene leads to the assembly of icosahedral core-less particles. *J. Virol.* **76**(6), 2654-66.

Andres, G., Garcia-Escudero, R., Simon-Mateo, C., and Vinuela, E. (1998). African swine fever virus is enveloped by a two-membraned collapsed cisterna derived from the endoplasmic reticulum. *J. Virol.* **72**(11), 8988-9001.

Andres, G., Simon-Mateo, C., and Vinuela, E. (1993). Characterization of two African swine fever virus 220-kDa proteins: a precursor of the major structural protein p150 and an oligomer of phosphoprotein p32. *Virology* **194**(1), 284-93.

Andres, G., Simon-Mateo, C., and Vinuela, E. (1997). Assembly of African swine fever virus: role of polyprotein pp220. *J. Virol.* **71**(3), 2331-41.

Angulo, A., Alcami, A., Vinuela, E. (1993). Virus-host interactions in African swine fever: the attachment to cellular receptors. *Arch. Virol. Suppl.* **7**, 169-183.

Angulo, A., Vinuela, E., and Alcami, A. (1992). Comparison of the sequence of the gene encoding African swine fever virus attachment protein p12 from field virus isolates and viruses passaged in tissue culture. *J. Virol.* **66**(6), 3869-72.

Arenas, M., Valiente, G., Posada, D. (2008). Characterization of reticulate networks based on the coalescent with recombination. *Mol. Biol. Evol.* **25**, 2517-2520.

Arguello-Astorga, G.R., Ascencia-Ibanez, J.T., Dallas, M.M., Orozco, B.M. and Hanley-Bowdoin, L. (2007). High-frequency reversion of geminivirus replication protein mutants during infection. *J. Virol.* **81**, 11005-11015.

Arias, M., and Sánchez-Vizcaíno, J. M. (2002a). African swine fever. Trends in Emerging Viral Infections of Swine. Morilla, A. ; Yoon, K.-J. ; Zimmerman, J. (eds) Iowa State University Press, Ames, 119-124.

Arias, M., and Sánchez-Vizcaíno, J. M. (2002b). African swine fever eradication: The Spanish model. Trends in Emerging Viral Infections of Swine. Morilla, A. ; Yoon, K.-J. ; Zimmerman, J. (eds) Iowa State University Press, Ames, 133-139.

Aris-Brosou, S., Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51**, 703-714.

Arzuza, O., Urzainqui, A., Diaz-Ruiz, J. R., and Tabares, E. (1992). Morphogenesis of African swine fever virus in monkey kidney cells after reversible inhibition of replication by cycloheximide. *Arch. Virol.* **124**(3-4), 343-54.

Avery OT, M.C., McCarty, M. (1944). Studies on the chemical nature of the substance inducing transfromation of pneumococcal types: induction of transformation by desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* **79**, 137-158.

Baer, KEv (1827). *De Ovi Mammalium et Hominis Generi*. Leipzig.

Bandelt, H.J., Macaulay, V., Richards, M. (2000). Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phylogenet. Evol.* **16**, 8-28.

Bandelt, H. J. a. D., A. (1992a). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics* **92**, 47-105.

Bandelt, H. J., and Dress, A. W. (1992b). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1**(3), 242-52.

Barderas, M. G., Rodriguez, F., Gomez-Puertas, P., Aviles, M., Beitia, F., Alonso, C., and Escribano, J. M. (2001). Antigenic and immunogenic properties of a chimera of two immunodominant African swine fever virus proteins. *Arch. Virol.* **146**(9), 1681-91.

Barry, D., Hartigan, J.A. (1987). Asynchronous distance between homologous DNA sequences. *Biometrics* **43**, 261-276.

Bashford, D., Chothia, C., and Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**(1), 199-216.

Basto, A.P., Nix, R.J., Boinas, F., et al. (2006). Kinetics of African swine fever virus infection in Ornithodoros erraticus ticks. *J. Gen. Virol*. **87**, 1863-1871.

Bastos, A. D., Penrith, M. L., Cruciere, C., Edrich, J. L., Hutchings, G., Roger, F., Couacy-Hymann, E., and G, R. T. (2003). Genotyping field strains of African swine fever virus by partial p72 gene characterisation. *Arch. Virol.* **148**(4), 693-706.

Beadle, G.W., Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci. USA* **27**, 499-506.

Beck, I.A., Drennan, K.D., Melvin, A.J. *et al.* (2001). Simple, sensitive, and specific detection oh human immunodeficiency virus type 1 subtype B DNA in dried blood samples for diagnosis in infants in the fiel. *J. Clin. Microbiol*. **39**, 29-33.

Becker, Y. (1998). Molecular Evolution of Viruses - Past and Present, Part 2, an introduction. *Virus Genes* **16**, 7-11.

Bernard, H. U. (1994). Coevolution of papillomaviruses with human populations. *Trends Microbiol.* **2**(4), 140-3.

Bernardi, G. (1993a). The isochore organization of the human genome and its evolutionary history--a review. *Gene* **135**, 57-66.

Bernardi, G. (1993b). The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.* **10**, 186-204.

Blasco, R., Aguero, M., Almendral, J. M., and Vinuela, E. (1989a). Variable and constant regions in African swine fever virus DNA. *Virology* **168**(2), 330-8.

Blasco, R., de la Vega, I., Almazan, F., Aguero, M., and Vinuela, E. (1989b). Genetic variation of African swine fever virus: variable regions near the ends of the viral DNA. *Virology* **173**(1), 251-7.

Blench, R.M. (1999). A history of pigs in Africa. In: Blencj, R.M., MacDonald, K., editors. Origins and development of African lifestock: archeology, genetics, linguistics and ethnography. *Florence, K.Y.: Routledge Books*, pp. 335-367.

Bock, R., Timmis, J.N. (2008). Reconstructing evolution: gene transfer from plastids to the nucleus. *Bioessays* **30**, 556-566.

Bolnick, DIaB.M.F. (2007). Sympatric speciation: Models and empirical evidence. Annual Review of Ecology, Evolution and *Systematics* **38**, 459-487.

Bonhoeffer, S., Holmes, E. and Nowak, M.A. (1995). Causes of HIV diversity. *Nature* **418**, 144.

Borca, M. V., C., C., Zsak, L., Laereid, W. W., Kutish, G. F., Neilan, J. G., Burrage, T. G., and Rock, D. L. (1998). Deletion of a CD2-like gene, 8-DR, from African swine fever virus affects viral infection in domestic swine. *J. Virol.* **72**, 2881-2889.

Borca, M. V., Irusta, P., Carillo, C., Afonso, C. L., Burrage, T., and Rock, D. L. (1994). African swine fever virus structural protein P72 contains a conformational neutralizing epitope. *Virology* **201**, 413-418.

Borde, V. (2007). The multiple roles of the Mre11 complex for meiotic recombination. *Chromosome Res*. **15**, 551-563.

Boshoff, C. I., Bastos, A. D., Gerber, L. J., and Vosloo, W. (2007). Genetic characterisation of African swine fever viruses from outbreaks in southern Africa (1973-1999). *Vet. Microbiol.* **121**(1-2), 45-55.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La Scola, B., Koonin, E. V., and Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci. USA* **106**(51), 21848-53.

Bromham, L., Penny, D. (2003). The modern molecular clock. *Nat. Rev. Genet.* **4**, 216-224.

Bromham, L., Penny, D., Rambaut, A., Hendy, M.D. (2000). The power of relative rates tests depends on the data. *J. Mol. Evol.* **50**, 296-301.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. In: Janovitz, M.F., Lapointe, F.J., McMorris, F.R., Mirkin, B., Roberts, F.S (Eds.), Bioconsensus, American Mathematical Society Publications, Piscataway, NJ, pp. 163-183.

Bryant, D., Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255-265.

Bubanovic, I., Najman, S., and Andjelkovic, Z. (2005). Origin and evolution of viruses: escaped DNA/RNA sequences as evolutionary accelerators and natural biological weapons. *Med. Hypotheses* **65**(5), 868-72.

Buendia, P., Narasimhan, G. (2009). Serial evolutionary networks of within-patient HIV-1 sequences reveal patterns of evolution of X4 strains. *BMC Syst. Biol.* **3**, 62.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. Mathematics in the Archeological and Historical Sciences, ed. F.R. Hodson, D.G. Kendall and P. Tautu. Edinburgh, UK: Edinburgh University Press, pp. 387-395.

Caride, E. e. a. (2002). Sexual transmission of HIV-1 isolate showing G-A hypermutation. *J. Clin. Virol.* **23**, 179-189.

Carrasco, L., Fernandez, A., Gomez Villamandos, J. C., Mozos, E., Mendez, A., and Jover, A. (1992). Kupffer cells and PIMs in acute experimental African swine fever. *Histol. Histopathol.* **7**(3), 421-5.

Carrascosa, A. L., del Val, M., Santaren, J. F., and Vinuela, E. (1985). Purification and properties of African swine fever virus. *J. Virol.* **54**(2), 337-44.

Carrascosa, J. L., Carazo, J. M., Carrascosa, A. L., Garcia, N., Santisteban, A., and Vinuela, E. (1984). General morphology and capsid fine structure of African swine fever virus particles. *Virology* **132**(1), 160-72.

Cavalli-Sforza, L.L., Edwards, A.W. (1967). Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233-257.

Chapman, D. A., Tcherepanov, V., Upton, C., and Dixon, L. K. (2008). Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. *J. Gen. Virol.* **89**(Pt 2), 397-408.

Chargaff, E., Zamenhof, S., Green, C. (1950). Composition of human desoxypentose nucleic acid. *Nature* **165**, 756-757.

Cheng, X.F., Wu, X.Y., Wang, H.Z., et al. (2012). High codon adaptation in citrus tristeza virus to its citrus host. *J. Virol*. **9**, 113.

Clark, A. G., Eisen, M. B., Smith, D. R., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167), 203-18.

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**(10), 1657-9.

Clewell, D.B. (2011). Tales of conjugation and sex pheromones: A plasmid and enterococcal odyssey. *Mob. Genet. Elements* **1**, 38-54.

Cobbold, C., Windsor, M., and Wileman, T. (2001). A virally encoded chaperone specialized for folding of the major capsid protein of African swine fever virus. *J. Virol.* **75**(16), 7221-9.

Codoner, F.M., Cuevas, J.M., Sanchez-Navarro, J.A., Pallas, V., Elena, S.F. (2005). Molecular evolution of the plant virus family Bromoviridae based on RNA3-encoded proteins. *J. Mol. Evol.* **61**, 697-705.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881-10890.

Crick, F.H., Griffith, J.S., Orgel, L.E. (1957). Codes without Commas. *Proc. Natl. Acad. Sci. USA* **43**, 416-421.

Cunha, C. V., and Costa, J. V. (1992). Induction of ribonucleotide reductase activity in cells infected with African swine fever virus. *Virology* **187**(1), 73-83.

Cutler, D.J. (2000). Estimating divergence times in the presence of an overdispersed molecular clock. *Mol. Biol. Evol.* **17**, 1647-1660.

De la Vega, I., Gonzalez, A., Blasto, R., Calvo, V., and and Viñuela, E. (1994). Nucleotide Sequence and Variability of the Inverted Terminal Repetitions of African Swine Fever Virus DNA. *Virology* **201**, 152-156.

De la Vega, I., Vinuela, E., and Blasco, R. (1990). Genetic variation and multigene families in African swine fever virus. *Virology* **179**(1), 234-46.

De Swart, R.L., Nur, Y., Abdallah, A., *et al*. (2001). Combination of reverse transcriptase PCR analysis and immunoglobulin M detection on filter paper blood samples allows diagnostic and epidemiological studies of measles. *J. Clin. Microbiol.* **39**, 270-273.

DeBoer, C. J. D., Hess, W. R., and Dardiri, A. H. (1969). Studies to determine the presence of neutralizing antibody in sera and kidney from swine recovered from African swine fever. Arch. *Gesamte Virusforsch* **27**, 44-54.

Diel, D.G., de Silva, L.H., Liu, H. *et al*. (2012). Genitic diversity of avian paramyxovirus type 1: proposal for unified nomenclature and classification system of Newcastle disease virus genotypes. Infect. Genet. Evol. **12**(8), 1770-1779.

Dixon, L.K., Escribano, J.M., Martins, C., Rock, D.L., Salas, M.L., and and Wilkinson, P.J. (2005). In: Fauquet, C.M., Mayo, M.A., Maniloff, J., Desselbelger, U., Ball, L.A. (Eds), Virus Taxonomy. VIII. Report of the ICTV. Elsevier/Academic Press, London, pp. 135-143.

Dixon, L.K., Twigg, S.R., Baylis, S.A., Vydelingum, S., Bristow, C., Hammond, J.M., and Smith, G.L. (1994). Nucleotide sequence of a 55 kbp region from the right end of the genome of a pathogenic African swine fever virus isolate (Malawi LIL20/1). *J. Gen. Virol.* **75**(Pt7), 1655-84.

Dixon, L.K., Wilkinson, P.J. (1988). Genetic diversity of African swine fever virus isolates from soft ticks (Ornithodoros moubata) inhaviting warthogs burrows in Zambia. *J. Gen. Virol.* **69** (Pt 12), 2981-2993.

Dobzhansky, T. (1937). Genetics and the Origin of Species. Columbia University Press.

Domingo, E. (1997). RNA virus evolution, population dynamics, and nutritional status. *Biol. Trace Elem. Res.* **56**(1), 23-30.

Domingo, E., and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* **51**, 151-78.

Doron-Faigenboim, A., Pupko, T. (2007). A combined empirical and mechanistic codon model. *Mol. Biol. Evol.* **24**, 388-397.

Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA* **90**(9), 4171-5.

Drake, J. W. A. (1991). A constant rate od spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA* **88**, 7160-7164.

Drake, J. W., and Hwang, C. B. (2005). On the mutation rate of herpes simplex virus type 1. Genetics **170**(2), 969-70.

Drake, J. W., Bebenek, A., Kissling, G. E., and Peddada, S. (2005). Clusters of mutations from transient hypermutability. *Proc. Natl. Acad. Sci. USA* **102**(36), 12849-54.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* **148**(4), 1667-86.

Drummond, A., Pybus, O.G., Rambaut, A. (2003a). Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* **54**, 331-358.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214.

Drummond, A.J., Ho, S.Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetic and dating with confidence. *PLoS Biol.* **4**(5), e88.

Drummond, A.J., Pybus, O.G.,Rambaut, A.,Forsberg, R. and Rodrigo, A.G. (2003b). Measurably evolving populations. *TRENDS in Ecology and Evolution* **18**, 481-488.

Dubay, S. A., Rosenstock, S. S., Stallknecht, D. E., deVos, J. C., Jr. (2006). Determining prevalence of bluetongue and epizootic hemorrhagic disease viruses in mule deer in Arizona (USA) using whole blood dried on paper strips compared to serum analyses. *J. Wildl. Dis*. **42**(1), 159-163.

Duffy, S., and Holmes, E.C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J. Virol.* **82**(2), 957-65.

Echols, H., Goodman, M.F. (1991). Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* **60**, 477-511.

Edgar, R.C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.

Eigen, M. (2002). Error catastrophe and antiviral strategy. *Proc. Natl. Acad. Sci. USA* **99**(21), 13374-6.

Eigen, M., Schuster, P. (1979). The Hypercycle ; A Principe of Natural Self-Organization. Berlin: Springer.

FAO (2002). RECONNAITRE LA PESTE PORCINE AFRICAINE. Manuel FAO de santé animale, ISBN 92-5-204471-X.

Faris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**, 645-668.

Felsenstein, J. (1973a). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471-492.

Felsenstein, J. (1973b). Maximum likelihood and minimum- steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**, 240-249.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annual Review of *Genetics* **22**, 521-565.

Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Felsenstein, J., and Churchill, G.A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93-104.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**(6), 368-76.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791.

Fernandez, A., Perez, J., Carrasco, L., Bautista, M. J., Sanchez-Vizcaino, J. M., and Sierra, M. A. (1992a). Distribution of ASFV antigens in pig tissues experimentally infected with two different Spanish virus isolates. *Zentralbl Veterinarmed B* **39**(6), 393-402.

Fernandez, A., Perez, J., Carrasco, L., Sierra, M. A., Sanchez-Vizcaino, M., and Jover, A. (1992b). Detection of African swine fever viral antigens in paraffin-embedded tissues by use of immunohistologic methods and polyclonal antibodies. *Am. J. Vet. Res.* **53**(8), 1462-7.

Filee J., Chandler, M. (2010). Gene exchange and the origin of giant viruses. *InterVirology* **53**, 354-361.

Filee, J., Forterre, P., Laurent, J. (2003). The role played by viruses in the evolution of their hosts: a review based on informational protein phylogénies. *Res. Microbiol.* **154**, 237-243.

Firth, C., Charleston, M. A., Duffy, S., Shapiro, B., and Holmes, E. C. (2009). Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J. Virol.* **83**(24), 12813-21.

Firth, C., Kitchen, A., Shapiro, B., *et al*. (2010). Using time-structured data to estimate evolutionary rates of double-strand DNA viruses. *Mol. Biol. Evol*. **27**, 2038-2051.

Fisher, R. A. (1930). The Genetical Theory of Ntural Selection. UK: Clarendon Press.

Fitch, W.M. (1967). Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.* **26**, 499-507.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113.

Fitch, W.M. (1971). Toward defining the course of evolution : Minimal change for a specific tree topology. *Systematic Zoology* **20**, 406-416.

Flint, S. J., Enquist, L.W., Racaniello, V.R and Skalka, A.M. (2004). Pinciples of *Virology* Molecular Biology. *Pathogenesis and Control of Animal Viruses*, ASM, Washington.

Foerstner, K.U., von Mering, C., Hooper, S.D., Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep* **6**, 1208-1213.

Forterre, P. (2002). The origine of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525-532.

Forterre, P. (2005). The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie* **87**, 793-803.

Forterre, P. (2006). Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc. Natl. Acad. Sci. USA* **103**(10), 3669-74.

Forterre, P., and Gadelle, D. (2009). Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. *Nucleic Acids Res.* **37**(3), 679-92.

Forterre, P., Filee, J., Myllykallio, H. (2004). Origin and evolution of DNA and DNA replication machineries. In: L. Ribas de pouplana (Ed.), The Genetic Code and the Origin of Life, *Landes Bioscience*, 145-168.

French, R. a. S., D.C. (2003). Evolution of wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. *Annu. rev. Phytopathol.* **41**, 199-214.

Friedrich, T.C., Dodds, E.J., Yant, L.J., et al. (2004). Reversion of CTL escape-variant immunodeficiency viruses in vivo. *Nat. Med.* **10**, 275-281.

Froissart, R., Roze, D., Uzest, M., *et al*. (2005). Recombination every day : abundant recombination in a virus during a single multi-cellular host infection. *Plos Blio*. **3**, e89.

Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., and Nishikawa, K. (2003). Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Evol.* **327**, 347-357.

Gallardo, C., Anchuelo, R., Pelayo, V., Poudevigne, F., Leon, T., Nzoussi, J., Bishop, R., Perez, C., Soler, A., Nieto, R., Martin, H., and Arias, M. (2011a). African swine fever virus p72 genotype IX in domestic pigs, Congo, 2009. *Emerg. Infect. Dis.* **17**(8), 1556-8.

Gallardo, C., Mwaengo, D. M., Macharia, J. M., Arias, M., Taracha, E. A., Soler, A., Okoth, E., Martin, E., Kasiti, J., and Bishop, R. P. (2009). Enhanced discrimination of African swine fever

virus isolates through nucleotide sequencing of the p54, p72, and pB602L (CVR) genes. *Virus Genes* **38**(1), 85-95.

Gallardo, C., Okoth, E., Pelayo, V., Anchuelo, R., Martin, E., Simon, A., Llorente, A., Nieto, R., Soler, A., Martin, R., Arias, M., and Bishop, R. P. (2011b). African swine fever viruses with two different genotypes, both of which occur in domestic pigs, are associated with ticks and adult warthogs, respectively, at a single geographical site. *J. Gen. Virol.* **92**(Pt 2), 432-44.

Galtier, N., Gouy, M., and Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl. Biosci.* **12**(6), 543-8.

Garcia-Arenal, F., Fraile, A., and Malpica, J. M. (2001). Variability and genetic structure of plant virus populations. *Annu. Rev. Phytopathol.* **39**, 157-86.

Garcia-Beato, R., Salas, M. L., Vinuela, E., and Salas, J. (1992). Role of the host cell nucleus in the replication of African swine fever virus DNA. *Virology* **188**(2), 637-49.

Garcia-Diaz, M. a. B., K. (2007). Multiple functions of DNA polymerases. *Crit. Rev. Plant. Sci.* **26**, 105-122.

Garcia-Escudero, R., Andres, G., Almazan, F., and Vinuela, E. (1998). Inducible gene expression from African swine fever virus recombinants: analysis of the major capsid protein p72. *J. Virol.* **72**(4), 3185-95.

Garcia-Escudero, R., Garcia-Diaz, M., Salas, M. L., Blanco, L., Salas, J. (2003). DNA polymerase X of African swine fever virus: insertion fidelity of gapped DNA substrate and AP lyase activity support a role in base excision repair of viral DNA. *J. Mol. Biol*. **326**, 1403-1412.

Giammarioli, M., Gallardo, C., Oggiano, A., Iscaro, C., Nieto, R., Pellegrini, C., Dei Giudici, S., Arias, M., and De Mia, G. M. (2011). Genetic characterisation of African swine fever viruses from recent and historical outbreaks in Sardinia (1978-2009). *Virus Genes* **42**(3), 377-87.

Gibbs, A. J., Calisher, C.H. and Garcia-Arenal, F. (1995). Molecular basis of virus evolution. Cambridge University Press, New York.

Gibbs, M. J., Armstrong, J. S., and Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**(7), 573-82.

Gilbert, W. (1986). The RNA world. *Nature* **319**, 618.

Gilford-Gonzales DaH, O. (2001). Domesticating animals in Africa: implications of Genetics and Archeological Findings. *J. World Prehist*. **24**, 1-23.

Goatley, L. C., Twigg, S. R., Miskin, J. E., Monaghan, P., St-Arnaud, R., Smith, G. L., and Dixon, L. K. (2002). The African swine fever virus protein j4R binds to the alpha chain of nascent polypeptide-associated complex. *J. Virol.* **76**(19), 9991-9.

Gojobori, T., Li, W.H., Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360-369.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182-198.

Goldman, N., Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725-736.

Gómez-Puertas, P., Rodriguez, F., Oviedo, J. M., Brun, A., Alonso, C., and Escribano, J. M. (1998). The African swine fever virus proteins p54 and p30 are involved in two distinct steps of virus attachment and both contribute to the antibody-mediated protective immune response. *Virology* **243**, 461-471.

Gomez-Puertas, P., Rodriguez, F., Oviedo, J. M., Ramiro-Ibanez, F., Ruiz-Gonzalvo, F., Alonso, C., and Escribano, J. M. (1996). Neutralizing antibodies to different proteins of African swine fever virus inhibit both virus attachment and internalization. *J. Virol.* **70**(8), 5689-94.

Gonzague, M., Roger, F., Bastos, A., Burger, C., Randriamparany, T., Smondack, S., and Cruciere, C. (2001). Isolation of a non-haemadsorbing, non-cytopathic strain of African swine fever virus in Madagascar. *Epidemiol. Infect.* **126**(3), 453-9.

Gonzalez, A., Calvo, V., Almazan, F., Almendral, J. M., Ramirez, J. C., de la Vega, I., Blasco, R., and Vinuela, E. (1990). Multigene families in African swine fever virus: family 360. *J. Virol.* **64**(5), 2073-81.

Gonzalez, A., Talavera, A., Almendral, J. M., and Vinuela, E. (1986). Hairpin loop structure of African swine fever virus DNA. *Nucleic Acids Res.* **14**(17), 6835-44.

Gouy, M., Guindon, S., Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221-224.

Grantham, R., Gautier, C., Gouy, M. (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* **8**, 1893-1912.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**, r43-74.

Greig, A. (1972). The localization of African swine fever virus in the tick Ornithodoros moubata porcinus. *Arch Gesamte Virusforsch* **39**, 240-247.

Grishin, N.V. (1997). Estimation of evolutionary distances from protein spatial structures. *J. Mol. Evol.* **45**, 359-369.

Haldane, J.B.S. (1932). The Causes of Evolution. Longman, Green and Co., Princeton University Press.

Hamers, R.L., Sigaloff, K.C., Wensing, A.M., et al. (2012). Patterns of HIV-1 drug resistance after first-line antiretroviral therapy (ART) failure in 6 sub-Saharan African countries: implications for second-line ART strategies. *Clin. Infect. Dis.* **54**, 1660-1669.

Hanada, K., Suzuki, Y., Gojobori, T. (2004). A large variation in the rates of synonymous substitution in RNA viruses and its relationships to a diversity of virla infection and transmission modes. *Mol. Biol. Evol*. **21**, 1074-1080.

Haresnape, J. M. (1984). African swine fever in Malawi. *Trop Anim Health Prod* **16**, 123-125.

Haresnape, J. M., Wilkinson, P. J., and Mellor, P. S. (1988). Isolation of African swine fever virus from ticks of the Ornithodoros moubata complex (Ixodoidea: Argasidae) collected within the African swine fever enzootic area of Malawi. *Epidemiol. Infect.* **101**(1), 173-85.

Harvey, P.H., Pagel, M.D. (1991). The comparative method in evolutionary biology. Oxford Studies in Ecology and Evolution (R.M. May and P.H. Harvey, eds.) Oxford University Press, Oxford.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**(2), 160-74.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97-109.

Hatwell, J. N., and Sharp, P. M. (2000). Evolution of human polyomavirus JC. *J. Gen. Virol.* **81**(Pt 5), 1191-200.

Heath, C. M., Windsor, M., and Wileman, T. (2001). Aggresomes resemble sites specialized for virus assembly. *J. Cel.l Biol.* **153**(3), 449-55.

Heath, L., van der Walt, E., Varsani, A., and Martin, D. P. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J. Virol.* **80**(23), 11827-32.

Hennig, W. (1966). Phylogenetic systematics. University of Illinois Press, Urbana.

Henning, W. (1950). Grundzüge einer Theorie der phylogenetischen Systematik (Fondements d'une théorie de la systématique phylogénétique). Deutscher Zentralverlag, Berlin.

Hernaez, B., Escribano, J. M., and Alonso, C. (2008). African swine fever virus protein p30 interaction with heterogeneous nuclear ribonucleoprotein K (hnRNP-K) during infection. FEBS **Lett 582**(23-24), 3275-80.

Hess, W. R. (1981). African swine fever: A Reassessment. *Advances in veterinary science and comparative medicine* **25**, 39-69.

Hess, W.R., Endris, R.G., Lousa, A., Caiado, J.M. (1989). Clearance of African swine fever from infected ticks (Acari) colonies. *J. Med. Entomol*. **26**, 314-317.

Higgins, D.G., Thompson, J.D., Gibson,T.J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383-402.

Holder, M., and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**(4), 275-84.

Holmes, E.C. (2003). Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77**(7), 3893-7.

Holmes, E.C. (2004). The phylogeography of human viruses. *Mol. Ecol*. **13**, 745-756.

Holmes, E.C., Drummond, A.J. (2007). The evolutionary genetics of viral emergence. *Curr. Top. Microbiol. Immunol*. **315**, 51-66.

Huelsenbeck, J.P., Helllis, D.M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology* **42**, 247-264.

Huelsenbeck JPaB, J.P. (2001). Application of the likelihood function in phylogenetic analysis. In : D.J. Balding, M. Bishop & C. Cannings (eds.), Handbook of Statistical Genetics, John Wiley and Sons, Inc., New York, pp. 415-439.

Huelsenbeck, J.P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8), 754-5.

Hughes, A.L., Irausquin, S., Friedman, R. (2009). The evolutionary biology of poxviruses. *Infect. Genet. Evol*. **10**, 50-59.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**(2), 254-67.

Huxley, J.S. (1942). Evolution : The Modern Synthesis. *Allen and Unwin*.

Ian J. Wilson, M. E. W., David J. Balding (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166**(2), 155-188.

Ingvarsson, P.K. (2008). Molecular evolution of synonymous codon usage in Populus. *BMC Evol. Biol.* **8**, 307.

Isnard, M., Granier, M., Frutos, R., Reynaud, B., and Peterschmitt, M. (1998). Quasispecies of three maize streak virus isolates obtained through different modes of selection from a population used to assess response to infection of maize cultivars. *J. Gen. Virol.* **79** (Pt 12), 3091-9.

Iyer, L. M., Aravind, L., and Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **75**(23), 11720-34.

Iyer, L. M., Balaji, S., Koonin, E. V., and Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* **117**(1), 156-84.

Jeffares, D. C., Poole, A. M., and Penny, D. (1998). Relics from the RNA world. *J. Mol. Evol.* **46**(1), 18-36.

Jenkins, G. M., Rambaut, A., Pybus, O. G., and Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* **54**(2), 156-65.

Jobb, G., von Haeseler, A., and Strimmer, K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* **4**, 18.

Jukes THaC, C.R. (1969). Evolution of protein molecules. In Mammalain protein metabolism, Munro, H.D. (ed.), Academic Press, New York, 21-132.

Katz, R. S., Premenko-Lanier, M., McChesney, M. B., Rota, P. A., Bellini, W. J. (2002). Detection of measles virus RNA in whole blood stored on filter paper. *J. Med. Virol.* **67**(4), 596-602.

Kidd, K.K. and Sgaramella-Zonta, L.A. (1971). Phylogentic analysis: concepts and methods. American Journal of Human *Genetics* **23**, 235-252.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624-626.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275-276.

Kimura, M.(1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

King, R.C. S.W.D., Mulligan, P.K. (2006). A dictionary of genetics. 7th ed. Oxford, 44.

King, K., Chapman, D., Argilaguet, J. M., Fishbourne, E., Hutet, E., Cariolet, R., Hutchings, G., Oura, C. A., Netherton, C. L., Moffat, K., Taylor, G., Le Potier, M. F., Dixon, L. K., Takamatsu, H. H. (2011). Protection of European domestic pigs from virulent African isolates of African swine fever virus by experimental immunisation. *Vaccine* **29**(28), 4593-600.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**, 235-248.

Kirkegaard, K., Baltimore, D. (1986). The mechanism of RNA recombination in poliovirus. *Cell* **47**, 433-443.

Kishino, H., Thorne, J. L., and Bruno, W. J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**(3), 352-61.

Kleiboeker, S.B., Scoles, G.A. (2001). Pathogenesis of African swine fever in Ornithodoros ticks. *Anim. Health Res. Rev.* **2**, 121-128.

Koonin, E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Micorbiol.* **1**, 127-136.

Koonin, E.V. and Yutin, N. (2010). Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Inter*Virology* **53**(5), 284-92.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29.

Koshi, J.M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* **42**(2), 313-20.

Kosiol, C., Holmes, I., Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* **24**, 1464-1479.

Kuhner, M.K., Yamato, J., Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421-1430.

Kuhner, M.K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**(6), 768-70.

Kumar, S., Hedges, S.B. (1998). A molecular timescale for vertebrate evolution. *Nature* **392**, 917-920.

Kusimba CMaK, S.B. (2000). Hinterlands and cities : Archeological investigations of economy and trade in Tsavo, out-eastern Kenya. *Department of Anthropology, The field Museum, 1400 S. Lake Shore Drive, Chigogo, Illinois, USA, 606005* **54**, 13-24.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., and Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* **455**(7209), 100-4.

Lamarche, B.J., Showalter, A.K., Tsai, M.D. (2005). An error-prone viral DNA ligase. *Biochemistry* **44**, 8408-8417.

Lamarche, B.J., Tsai, M.D. (2006). Contributions of an endonuclease IV homologue to DNA repair in the African swine fever virus. *Biochemistry* **45**, 2790-2803.

Lamarche, B.J., Kumar, S., and Tsai, M.-D. (2006). ASFV DNA Polymerase X is Extremely Error-Prone Under Diverse Assay Conditions and Within Multiple DNA Sequence Contexts. *Biochemistry* **45**(49), 14826-14833.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**(1), 86-93.

Lancaster, K. Z., and Pfeiffer, J. K. (2011). Mechanisms controlling virulence thresholds of mixed viral populations. *J. Virol.* **85**(19), 9778-88.

Lancaster, K. Z., and Pfeiffer, J. K. (2012). Viral population dynamics and virulence thresholds. *Curr. Opin. Microbiol.* **15**(4), 525-30.

Legendre, P., Makarenkov, V. (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.* **51**, 199-216.

Leitao, A., Cartaxeiro, C., Coelho, R., Cruz, B., Parkhouse, R. M., Portugal, F., Vigario, J. D., and Martins, C. L. (2001). The non-haemadsorbing African swine fever virus isolate ASFV/NH/P68 provides a model for defining the protective anti-virus immune response. *J. Gen. Virol.* **82**(Pt 3), 513-23.

Leitao, A., Malur, A., Cornelis, P., and Martins, C. L. (1998). Identification of a 25-aminoacid sequence from the major African swine fever virus structural protein VP72 recognised by porcine cytotoxic T lymphocytes using a lipoprotein based expression system. *J. Virol. Methods* **75**(1), 113-9.

Levathes, L.E. (1994). When China Ruled the Seas: The Treasure Fleet of the Dragon Throne, 1405-1433. *New York: Oxford University Press*.

Li, B., Gladden, A.D., Altfeld, M., *et al*. (2007). Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J. Virol.* **81**, 193-201.

Li, W. H. (1997). Molecular Evolution. Sinauer Associates, Sunderland, Massachusetts.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**(11), 1451-2.

Lobry, JRaN, A. (2006). Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* **30**, 128–136.

Lockart, P.J., Steel, M., Hendy, M.D. and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of evolution. *Molecular Biology and Evolution* **11**, 605-612.

Lopez-Otin, C., Simon-Mateo, C., Martinez, L., and Vinuela, E. (1989). Gly-Gly-X, a novel consensus sequence for the proteolytic processing of viral and cellular proteins. *J. Biol. Chem.* **264**(16), 9107-10.

Lubisi, B. A., Bastos, A. D., Dwarka, R. M., and Vosloo, W. (2005). Molecular epidemiology of African swine fever in East Africa. *Arch. Virol.* **150**(12), 2439-52.

Lubisi, B. A., Bastos, A. D., Dwarka, R. M., and Vosloo, W. (2007). Intra-genotypic resolution of African swine fever viruses from an East African domestic pig cycle: a combined p72-CVR approach. *Virus Genes* **35**(3), 729-35.

Lubisi, B. A., Dwarka, R. M., Meenowa, D., and Jaumally, R. (2009). An investigation into the first outbreak of African swine fever in the Republic of Mauritius. *Transbound Emerg. Dis.* **56**(5), 178-88.

Lynch, M., and Crease, T. J. (1990). The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* **7**(4), 377-94.

Ma, M.R., Ha, X.Q., Ling, H., et al. (2011). The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. *J. Virol.* **8**, 544.

Makarenkov, V., Legendre, P. (2004). From a phylogenetic tree to a reticulated network. J *Comput. Biol.* **11**, 195-212.

Malogolovkin, A., Yelsukova, A., Gallardo, C., Tsybanov, S., and Kolbasov, D. (2012). Molecular characterization of African swine fever virus isolates originating from outbreaks in the Russian Federation between 2007 and 2011. *Vet. Microbiol.* **158**(3-4), 415-9.

Martin, D., and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**(6), 562-3.

Martin Hernandez, A. M. and Tabares, E. (1991). Expression and characterization of the thymidine kinase gene of African swine fever virus. *J. Virol.* **65**(2), 1046-52.

Martins, C. L. V., and Leitao, A. C. (1994). Porcine immune responses to African swine fever virus (ASFV) infection. Veterinary Immunology and Immunopathology **43**, 99-106.

Martins, C. L., Lawman, M. J., Scholl, T., Mebus, C. A., and Lunney, J. K. (1993). African swine fever virus specific porcine cytotoxic T cell activity. *Arch. Virol.* **129**(1-4), 211-25.

Masel, J. (2011). Genetic drift. *Curr. Biol.* **20**, R837–R838.

Mbayed, V. A., Lopez, J. L., Telenta, P. F., Palacios, G., Badia, I., Ferro, A., Galoppo, C., and Campos, R. (1998). Distribution of hepatitis B virus genotypes in two different pediatric populations from Argentina. *J. Clin. Microbiol.* **36**(11), 3362-5.

McCormack, J. E., Huang, H., Knowles, L. L. (2009). Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. Biol.* **58**(5), 501-508.

McCullough, K.C., Basta, S., Knotig, S., et al. (1999). Intermediate stages in monocyte-macrophage differentiation modulate phenotype and susceptibility to virus infection. *Immunology* **98**, 203-212.

McGeoch, D. J., and Gatherer, D. (2005). Integrating reptilian herpesviruses into the family herpesviridae. *J. Virol.* **79**(2), 725-31.

McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E., and Telford, E. A. (1995). Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J. Mol. Biol.* **247**(3), 443-58.

McGeoch, D. J., Dolan, A., and Ralph, A. C. (2000). Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74**(22), 10401-6.

McGeoch, D. J., Gatherer, D., and Dolan, A. (2005). On phylogenetic relationships among major lineages of the Gammaherpesvirinae. *J. Gen. Virol.* **86**(Pt 2), 307-16.

Mebus, C.A., McVicar, J.W., Dardiri, A.H. (1981). Comparison of the pathology of high and low virulence African swine fever virus infections. In: Wilkinson P J, editor. Proceedings of CEC/FAO expert consultation in African swine fever research, Sardinia, Italy, September 1981. Luxemburg, Belgium: Commission of the European Communities ; 1981. pp. 183–194.

Meiering, C.D., and Linial, M. L. (2001). Historical perspective of foamy virus epidemiology and infection. *Clin. Microbiol. Rev.* **14**(1), 165-76.

Mendel, J.G. (1866). Versuche ueber Pflanzenhybriden : zwei Abhandlungen Verhandlungen des naturforschenden Vereines in Brünn tome IV, 3-47.

Metropolis, N., Ulam, S. (1949). The Monte Carlo method. *J. Am. Stat. Assoc.* **44**, 335-341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of chemical physics* **21**, 1087-1092.

Michelotti, E.F., Michelotti, G.A., Aronsohn, A.I., Levens, D. (1996). Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Mol. Cell. Biol.* **16**, 2350-2360.

Miller, R.V. (2001). Environmental bacteriophage-host interactions: factors contribution to natural transduction. *Antonie Van Leeuwenhoek* **79**, 141-147.

Miller, R.H. and Robinson, W. S. (1986). Common evolutionary origin of hepatitis B virus and retroviruses. *Proc. Natl. Acad. Sci. USA* **83**(8), 2531-5.

Minguez, I., Rueda, A., Dominguez, J., and Sanchez-Vizcaino, J. M. (1988). Double labeling immunohistological study of African swine fever virus-infected spleen and lymph nodes. *Vet. Pathol.* **25**(3), 193-8.

Misinzo, G., Magambo, J., Masambu, J., Yongolo, M. G., Van Doorsselaere, J. and Nauwynck, H. J. (2010). Genetic characterization of African swine fever viruses from a 2008 outbreak in Tanzania. *Transbound Emerg. Dis.* **58**(1), 86-92.

Montgomery, R. (1921). On a form of swine fever occurring in British East Africa (Kenya colony). *J. Comp. Pathol.* **34**, 159, 191, 243-262.

Moran, N.A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583-586.

Morrison, D.A. (2005). Networks in phylogenetic analysis: new tools for population biology. *Int J Parasitol* **35**, 567-582.

Mosquera Mdel, M., Echevarria, J. E., Puente, S., Lahulla, F., de Ory, F. (2004). Use of whole blood dried on filter paper for detection and genotyping of measles virus. *J. Virol. Methods*. **117**(1), 97-99.

Moulton, J., and Coggins, L. (1968). Comparison of lesions in acute and chronic African swine fever. *Cornell Vet* **58**(3), 364-88.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposium in Quantitative Biology* **51**, 263-273.

Mullis, K.B., Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335-350.

Muse, S.V., Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715-724.

Nakhleh, L., Sun, J., Warnow, T., et al. (2003). Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. *Pac. Symp. Biocomput.*, 315-326.

Needleman, SBaW, C.D. (1970). A general method applicable to the search for similariries in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

Nei, M. (1972). Genetic distances between populations. *American Naturalist* **106**, 283-292.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. . *Genetics* **76**, 379-390.

Nei, M. (2005). Bottlenecks, genetic polymorphism and speciation. *Genetics* **170**(1), 1-4.

Neilan, J. G., Zsak, L., Lu, Z., Burrage, T. G., Kutish, G. F., and Rock, D. L. (2004). Neutralizing antibodies to African swine fever virus proteins p30, p54, and p72 are not sufficient for antibody-mediated protection. *Virology* **319**(2), 337-42.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. *Statistical Decision Theory and Related Topics*, 1-27.

Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelop gene. *Genetics* **148**, 929-936.

Nix, R. J., Gallardo, C., Hutchings, G., Blanco, E., and Dixon, L. K. (2006). Molecular epidemiology of African swine fever virus studied by analysis of four variable genome regions. *Arch. Virol.* **151**(12), 2475-94.

Nogal, M. L., Gonzalez de Buitrago, G., Rodriguez, C., Cubelos, B., Carrascosa, A. L., Salas, M. L., and Revilla, Y. (2001). African swine fever virus IAP homologue inhibits caspase activation and promotes cell survival in mammalian cells. *J. Virol.* **75**(6), 2535-43.

Norley, S. G., and Wardley, R. C. (1983). Investigation of porcine natural-killer cell activity with reference to African swine-fever virus infection. *Immunology* **49**(4), 593-7.

Ochoa, G. (2006). Error thresholds in genetic algorithms. *Evol. Comput.* **14**(2), 157-82.

Ohno, S. (1970). Evolution by gene duplication. Springer-Verlag, New York.

Ojosnegros, S., Perales, C., Mas, A., and Domingo, E. (2011). Quasispecies as a matter of fact: viruses and beyond. *Virus Res.* **162**(1-2), 203-15.

Oliveros, M., Garcia-Escudero, R., Alejo, A., Vinuela, E., Salas, M. L., and Salas, J. (1999). African swine fever virus dUTPase is a highly specific enzyme required for efficient replication in swine macrophages. *J. Virol.* **73**(11), 8934-43.

Oliveros, M., Yanez, R. J., Salas, M. L., Salas, J., Vinuela, E., and Blanco, L. (1997). Characterization of an African swine fever virus 20-kDa DNA polymerase involved in DNA repair. *J. Biol. Chem.* **272**(49), 30899-910.

Onisk, D. V., Borca, M. V., Kutish, G., Kramer, E., Irusta, P., and Rock, D. L. (1994). Passively transferred African swine fever virus antibodies protect swine against lethal infection. *Virology* **198**(1), 350-4.

Osawa, S. (1995). Evolution of the genetic code. Oxford University Press, Oxford.

Owen, R.D. (1848). On the archetype and homologies of the vertebrate skeleton. John van Voorst, Londres.

Owolodun, O. A., Bastos, A. D., Antiabong, J. F., Ogedengbe, M. E., Ekong, P. S., and Yakubu, B. (2010). Molecular characterisation of African swine fever viruses from Nigeria (2003-2006) recovers multiple virus variants and reaffirms CVR epidemiological utility. *Virus Genes* **41**(3), 361-8.

Padidam, M., Sawyer, S. and Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**(2), 218-25.

Pan, I.C., Hess, W.R. (1985). Diversity of African swine fever. *Am. J. Vet. Res*. **46**, 314-320.

Pastor, M.J., Arias, M., Alcaraz, C., De Diego, M., Escribano, J.M. (1992). A sensitive dot immunobinding assay for serodiagnosos of African swine fever virus with application in field conditions. *J. Vet. Diag. Invest*. **4**, 254-257.

Pathak, V. K. a. T., H.M. (1992). 5-azacytidine and RNA secondary structure increase the retrovirus mutation rate. *J. Virol.* **66**, 3093-3100.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.* **51**, 41-47.

Pavesi, A. (2005). Utility of JC polyomavirus in tracing the pattern of human migrations dating to prehistoric times. *J. Gen. Virol.* **86**(Pt 5), 1315-26.

Pawlotsky, J.M. (2012). New antiviral agents for hepatitis C. *F1000 Biol. Rep.* **4**, 5.

Pedersen, A.K., Wiuf, C., Christiansen, F.B. (1998). A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**, 1069-1081.

Penrith, M. L., and Vosloo, W. (2009). Review of African swine fever: transmission, spread and control. *J. S. Afr. Vet. Assoc.* **80**(2), 58-62.

Penrith, M. L., Thomson, G. R., and Bastos, A. D. S. (2004). African swine fever. In: Coetzer, J.A.W., Tustin, R.C. (Eds.), Infectious Diseases of Livestock with Special Reference to Southern Africa, 2nd edi. Oxford University Press, Cape Town, 1087-1119.

Perales, C., Martin, V., Domingo, E. (2012). Lethal mutagenesis of viruses. *Curr. Opin. Virol.* **1**, 419-422.

Pereto, J., Lopez-Garcia, P., Moreira, D. (2004). Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem. Sci.* **29**, 469-477.

Petrosillo, N., Ippolito, G., Solforosi, L., Varaldo, P. E., Clementi, M., and Manzin, A. (2000). Molecular epidemiology of an outbreak of fulminant hepatitis B. *J. Clin. Microbiol.* **38**(8), 2975-81.

Plowright, W. (1977). Vector transmission of African swine fever virus. In "Seminar on Hog cholera, classical swine fever and African swine fever", pp. 575-587. Eur 5904EN, commission of the European communities.

Plowright, W. (1981). African swine fever. In: Davis, J.W., Karstad, L.H., Trainer, D.O., Editors. Infectious diseases of wild mammals. 2nd ed. Ames, Iowa Sate University Press., 178-190.

Plowright, W., Parker, J., and Peirce, M. A. (1969). African swine fever virus in ticks (Ornithodoros moubata, murray) collected from animal burrows in Tanzania. *Nature* **221**(5185), 1071-3.

Plowright, W., Perry, C. T., and Greig, A. (1974). Sexual transmission of African swine fever virus in the tick, Ornithodoros moubata porcinus, Walton. *Res. Vet. Sci.* **17**(1), 106-13.

Plowright, W., Perry, C. T., and Peirce, M. A. (1970a). Transovarial infection with African swine fever virus in the argasid tick, Ornithodoros moubata porcinus, Walton. *Res. Vet. Sci.* **11**(6), 582-4.

Plowright, W., Perry, C. T., Peirce, M. A., and Parker, J. (1970b). Experimental infection of the argasid tick, Ornithodoros moubata porcinus, with African swine fever virus. *Arch Gesamte Virusforsch* **31**(1), 33-50.

Plug laB, S. (2001). The distribution of macrommals in Southern Africa over the past 30,000 years. *Transvaal Museum Monograph*. **13**, South Africa.

Poole, A. M., Logan, D.T. (2005). Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA génome? *Mol. Biol. Evol.* **22**, 1444-1455.

Posada, D., and Crandall, K.A. (1998). ModelTest: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818.

Posada, D., Crandall, K.A. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* **16**, 37-45.

Powell, P. P., Dixon, L. K., and Parkhouse, R. M. E. (1996). An IκB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokines responses in host macrophages. *J. Virol.* **70**(12), 8527-8533.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395-399.

Rambaut, A., Bromham, L. (1998). Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* **15**, 442-448.

Ramirez, O., Ojeda, A., Tomazs, A. et al. (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol. Biol. Evol*. **26**, 2061-2072.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J. M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science* **306**(5700), 1344-50.

Ren, F., Tanaka, H., Yang, Z. (2005). An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* **54**, 808-818.

Revilla, Y., Callejo, M., Rodriguez, J. M., Culebras, E., Nogal, M. L., Salas, M. L., Viñuela, E., and Fresno, M. (1998). Inhibition of nuclear factor ?B activation by a virus-encoded I?B-like protein. *J. Biol. Chem.* **273**(9), 5405-5411.

Revilla, Y., Cebrian, A., Baixeras, E., Martinez, C., Viñuela, E., and Salas, M. L. (1997). Inhibition of apoptosis by African swine fever virus bcl-2 homologue: role of the BH1 domain. *Virology* **228**, 400-404.

Reynolds, J., Weir, B.S., Cockerham, C.C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767-779.

Rhodes, T., Wargo, H., Hu, W.S. (2003). High rates of human immunodeficiency virus type 1 recombination: near random segregation of markers one kilobase apart in one round of viral replication. *J. Virol*. **77**, 11193-11200.

Ribas de Pouplana, L. (2004). The Genetic Code and the Origin of Life. *Landes Bioscience*, **1**-253.

Rocha, E.P., Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291-294.

Rodriguez, F., Alcaraz, C., Eiras, A., Yanez, R. J., Rodriguez, J. M., Alonso, C., Rodriguez, J. F., and Escribano, J. M. (1994). Characterization and molecular basis of heterogeneity of the African swine fever virus envelope protein p54. *J. Virol.* **68**(11), 7244-52.

Rodriguez, F., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**(4), 485-501.

Rodriguez, J. M., Yanez, R. J., Almazan, F., Vinuela, E., and Rodriguez, J. F. (1993a). African swine fever virus encodes a CD2 homolog responsible for the adhesion of erythrocytes to infected cells. *J. Virol.* **67**(9), 5312-20.

Rodriguez, J. M., Yanez, R. J., Rodriguez, J. F., Vinuela, E., and Salas, M. L. (1993b). The DNA polymerase-encoding gene of African swine fever virus: sequence and transcriptional analysis. *Gene* **136**(1-2), 103-10.

Roger, F., Ratovonjato, J., Vola, P., and Uilenber, G. (2001). Ornithodoros porcinus ticks, bushpigs, and African swine fever in Madagascar. *Exp. Appl. Acarol.* **25**, 263-269.

Rojo, G., Chamorro, M., Salas, M. L., Vinuela, E., Cuezva, J. M., and Salas, J. (1998). Migration of mitochondria to viral assembly sites in African swine fever virus-infected cells. *J. Virol.* **72**(9), 7583-8.

Rojo, G., Garcia-Beato, R., Vinuela, E., Salas, M. L., and Salas, J. (1999). Replication of African swine fever virus DNA in infected cells. *Virology* **257**(2), 524-36.

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12), 1572-4.

Rouiller, I., Brookes, S. M., Hyatt, A. D., Windsor, M., and Wileman, T. (1998). African swine fever virus is wrapped by the endoplasmic reticulum. *J. Virol.* **72**(3), 2373-87.

Rowlands, R. J., Michaud, V., Heath, L., Hutchings, G., Oura, C., Vosloo, W., Dwarka, R., Onashvili, T., Albina, E., and Dixon, L. K. (2008). African swine fever virus isolate, Georgia, 2007. *Emerg. Infect. Dis.* **14**(12), 1870-4.

Ruiz Gonzalvo, F., Caballero, C., Martinez, J., and Carnero, M. E. (1986a). Neutralization of African swine fever virus by sera from African swine fever-resistant pigs. *Am. J. Vet. Res.* **47**(8), 1858-62.

Ruiz Gonzalvo, F., Carnero, M. E., Caballero, C., and Martinez, J. (1986b). Inhibition of African swine fever infection in the presence of immune sera in vivo and in vitro. *Am. J. Vet. Res.* **47**(6), 1249-52.

Ruiz-Gonzalvo, F., and Coll, J. M. (1993). Characterization of a soluble hemagglutinin induced in African swine fever virus-infected cells. *Virology* **196**(2), 769-77.

Ruiz-Gonzalvo, F., Rodriguez, F., and Escribano, J. M. (1996). Functional and immunological properties of the baculovirus-expressed hemagglutinin of African swine fever virus. *Virology* **218**(1), 285-9.

Rzhetsky, A., Nei, M. (1992a). Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* **35**, 367-375.

Rzhetsky, AaN, M. (1992b). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* **9**, 945-967.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**(4), 406-25.

Salas, M. L. (1999). African Swine Fever Virus (asfarviridae). Encyclopedia of Virology second edition 1, pp. 30-38.

Salemi, M., Lewis, M., Egan, J. F., Hall, W. W., Desmyter, J., and Vandamme, A. M. (1999). Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. *Proc. Natl. Acad. Sci. USA* **96**(23), 13253-8.

Sanchez, C., Domenech, N., Vazquez, J., Alonso, F., Ezquerra, A., and Dominguez, J. (1999). The porcine 2A10 antigen is homologous to human CD163 and related to macrophage differentiation. *J. Immunol.* **162**(9), 5230-7.

Sanchez-Torres, C., Gomez-Puertas, P., Gomez-del-Moral, M., Alonso, F., Escribano, J. M., Ezquerra, A., and Dominguez, J. (2003). Expression of porcine CD163 on monocytes/macrophages correlates with permissiveness to African swine fever infection. *Arch. Virol.* **148**(12), 2307-23.

Sanchez-Vizcaino, J. M. (2006). African swine fever. Diseases of Swine, 9th Edition. Blackwell Publishing . Chapter 13., pp291-298.

Sanderson, M.J. (1997). A non parametric approach to estimating divergence times in the absence of rate constancy. *J. Mol. Evol.* **14**, 1218-1231.

Sanger, F., Barrell, B.G., Brown, N.L ., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695.

Santti, J., Hyypia, T., Kinnunen, L., and Salminen, M. (1999). Evidence of recombination among enteroviruses. *J. Virol.* **73**(10), 8741-9.

Schierup, M.H. and Hein, J. (2000a). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**(2), 879-91.

Schierup, M.H. and Hein, J. (2000b). Recombination and the molecular clock. *Mol. Biol. Evol*. **17**, 1578-1579.

Schneider, A., Cannarozzi, G.M., Gonnet, G.H. (2005). Empirical codon substitution matrix. BMC *Bioinformatics* **6**, 134.

Scott, T.W., Weaver, S.C., Mallampalli, V.L. (1994). Evolution of mosquito-born viruses. Pp. 293-324 in S.S. Morse, ed. Evolutionary biology of viruses. Raven Press, New York.

Schwarz, G. (1978). Estimating the dimension of a model *Ann. Stat.* **6**, 461-464.

Sellers, P.H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26**, 787-793.

Shackelton, L. A., Parrish, C. R., Truyen, U., and Holmes, E. C. (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc. Natl. Acad. Sci. USA* **102**(2), 379-84.

Showalter, A.K., Byeon, I.J., Su, M.I., Tsai, M.D. (2001). Solution structure of a viral DNA polymerase X and evidence for a mutagenic function. *Nat. Struct. Biol.* **8**, 942-946.

Showalter, A.K., Tsai, M.D. (2001). A DNA polymerase with specificity for five base pairs. *J. Am. Chem. Soc.* **123**, 1776-1777.

Shriner, D., Nickle, D. C., Jensen, M. A., and Mullins, J. I. (2003). Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.* **81**(2), 115-21.

Sia, E.A., Kokoska, R.J., Dominska, M., Greenwel,l .P., Petes, T.D. (1997). Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.* **17**, 2851-2858.

Simon-Mateo, C., Andres, G., Almazan, F., and Vinuela, E. (1997). Proteolytic processing in African swine fever virus: evidence for a new structural polyprotein, pp62. *J. Virol.* **71**(8), 5799-804.

Simon-Mateo, C., Andres, G., and Vinuela, E. (1993). Polyprotein processing in African swine fever virus: a novel gene expression strategy for a DNA virus. *EMBO J.* **12**(7), 2977-87.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**(2), 126-9.

Sobol, R.W., Horton, J.K., Kuhn, R., et al. (1996). Requirement of mammalian DNA polymerase-beta in base-excision repair. *Nature* **379**, 183-186.

Strimmer, K., and Rambaut, A. (2002). Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.* **269**(1487), 137-42.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. Communication in Statist. *Theor. Meth.* **7**(1), 13-26.

Sumption, K. J., Hutchings, G. H., Wilkinson, P. J., and Dixon, L. K. (1990). Variable regions on the genome of Malawi isolates of African swine fever virus. *J. Gen. Virol.* **71** (Pt 10), 2331-40.

Sun, H., Jenson, J., Dixon, L. K., and Parkhouse, M. E. (1996). Characterization of the African swine fever virion protein j18L. *J. Gen. Virol.* **77** (Pt 5), 941-6.

Suzuki, Y., Katayama, K., Fukushi, S., Kageyama, T., Oya, A., Okamura, H., Tanaka, Y., Mizokami, M., and Gojobori, T. (1999). Slow evolutionary rate of GB virus C/hepatitis G virus. *J. Mol. Evol.* **48**(4), 383-9.

Swart, T., Kotze, A., Olivier, P.A.S. and Grobler J.P. (2010). Microsatellite-based characterization of Southern African domestic pigs (*sus scrofa domestica*). *South African Journal of Animal Science* **40**, 121-132.

Switzer, W. M., Salemi, M., Shanmugam, V., Gao, F., Cong, M. E., Kuiken, C., Bhullar, V., Beer, B. E., Vallet, D., Gautier-Hion, A., Tooze, Z., Villinger, F., Holmes, E. C., and Heneine, W. (2005). Ancient co-speciation of simian foamy viruses and primates. *Nature* **434**(7031), 376-80.

Tabares, E., Olivares, I., Santurde, G., Garcia, M. G., Martin, E., and Carmero, M. E. (1987). African swine fever virus DNA: deletions and additions during adaptation to growth in monkey kidney cells. *Arch. Virol.* **97**, 333-346.

Tajima, F. (1993). Measurement of DNA polymorphism. In Takahata, N. and Clark, A. G. (eds), Mechanisms of Molecular Evolution. Sinauer Associates. Inc., Sunderland, Massachusetts, 37-59.

Takezaki, N., Rzhetsky, A., Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823-833.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.* **9**, 678-687.

Tamura, K., Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512-526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetic analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**(10), 2731-9.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lecture Math. Life Sci.* **17**, 57-86.

Taylor, W.R. (1987). Multiple sequence alignment by a pairwise algorithm. *Comput Appl. Biosci.* **3**, 81-87.

Templeton, A.R., Crandall, K. A., and Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**(2), 619-33.

Thompson, J.D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**(22), 4673-80.

Thomson, G. R. (1985). The epidemiology of African swine fever: the role of free-living hosts in Africa. *Onderstepoort J. Vet. Res.* **52**(3), 201-9.

Thomson, G.R., Gainaru, M. D., and Van Dellen, A. F. (1980). Experimental infection of warthos (Phacochoerus aethiopicus) with African swine fever virus. *Onderstepoort J. Vet. Res.* **47**(1), 19-22.

Thorne, J.L., Kishino, H., Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647-1657.

Thorne, J.L., and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**(5), 689-702.

Tian, D., Wang, Q., Zhang, P., et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105-108.

Tidona, C. A., Schnitzler, P., Kehm, R., and Darai, G. (1998). Is the major capsid protein of iridoviruses a suitable target for the study of viral evolution? *Virus Genes* **16**(1), 59-66.

Tillier, E.R., Collins, R.A. (1998). High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**, 1993-2002.

Treanor, J. (2004). Influenza vaccine: outmaneuvering antigenic shift and drift. *N. Engl. J. Med.* **350**, 218-220.

Umar, A., Kunkel T.A. (1996). DNA-replication fidelity, mismatch repair and genome instability in cancer cells. *Eur. J. Biochem.* **238**, 297-307.

Valdeira, M.L., and Geraldes, A. (1985). Morphological study on the entry of African swine fever virus into cells. *Biol. Cell* **55**(1-2), 35-40.

Vali, U., Brandstrom, M., Johansson, M., Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet*. **9**, 8.

Vallee, I., Tait, S. W., and Powell, P. P. (2001). African swine fever virus infection of porcine aortic endothelial cells leads to inhibition of inflammatory responses, activation of the thrombotic state, and apoptosis. *J. Virol.* **75**(21), 10372-82.

Van de Peer, Y., Rensing, S., Maier, U.-G., De Watcher, R. (1996). Substitution rate calibration of small ribosomal subunit RNA identifies chlorarachnophytes endosumbionts as remnants of green algae. *Proceedings of the National Academy of Sciences, USA* **93**, 7732-7736.

Varsani, A., van der Walt, E., Heath, L. et al. (2006). Evidence of ancient papillomavirus recombination. *J. Gen. Virol*. **87**, 2527-2531.

Vigne, E., Marmonier A., Fuchs M. (2008). Multiple interspecies recombination events within RNA2 of Grapevine fanleaf virus and Arabis mosaic virus. *Arch. Virol.* **153**, 1771-1776.

Walsh, C. P. a. X., G.L. (2006). Cystosine methylation and DNA repair. *Curr. Topics Microbiol. Immunol.* **301**, 283-315.

Wang, C.Y., Giambrone, J.J., Smith, B.F. (2002). Detection of duck hepatitis B virus DNA on filter paper by PCR and SYBR green dye-based quantitative PCR. *J. Clin. Microbiol.* **40**, 2584-2590.

Wang, H.C., Susko, E., Roger, A.J. (2006). On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem. Biophys. Res. Commun.* **342**, 681-684.

Wang, W., Hellinga, H.W., Beese, L.S. (2011). Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc. Natl. Acad. Sci. USA* **108**, 17644-17648.

Wardley, R. C., and Wilkinson, P. J. (1985). An immunological approach to vaccines against African swine fever virus. *Vaccine* **3**(1), 54-6.

Wardley, R. C., de, M. A. C., Black, D. N., de Castro Portugal, F. L., Enjuanes, L., Hess, W. R., Mebus, C., Ordas, A., Rutili, D., Sanchez Vizcaino, J., Vigario, J. D., Wilkinson, P. J., Moura Nunes, J. F., and Thomson, G. (1983). African Swine Fever virus. Brief review. *Arch. Virol.* **76**(2), 73-90.

Wardley, R. C., Norley, S. G., Wilkinson, P. J., and Williams, S. (1985). The role of antibody in protection against African swine fever virus. *Vet. Immunol. Immunopathol.* **9**(3), 201-12.

Wardley, R. C., Wilkinson, P. J., and Hamilton, F. (1977). African swine fever virus replication in porcine lymphocytes. *J. Gen. Virol.* **37**(2), 425-7.

Warren, R. A. J. (1980). Modified bases in bacteriophage DNAs. *Annu. Rev. Microbiol.* **34**, 137-158.

Watson, J.D., Crick, F.H. (1953). Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738.

Weaver, S.C., and Reisen, W. K. (2009). Present and future arboviral threats. *Antiviral Res.* **85**(2), 328-45.

Wertheim, J.O., Pond, S.L.K. (2011). Purifying Selection Can Obscure the Ancient Age of Viral Lineages. *Mol. Biol. Evol*. **28**, 3355-3365.

Wesley, R.D., Quintero, J. C., and Mebus, C. A. (1984). Extraction of viral DNA from erythrocytes of swine with acute African swine fever. *Am. J. Vet. Res.* **45**(6), 1127-31.

Whyard, T.C., Wool, S. H., and Letchworth, G. J. (1985). Monoclonal antibodies against African swine fever viral antigens. *Virology* **142**(2), 416-20.

Wilkinson, P.J. (1989). Afican swine fever virus. In: "Virus Infections of Vertebrates. Vol. 2: Virus infections of porcines". (Ed. M.B. Penjaert.) pp. 17-35 (Elsevier).

Wilkinson, P.J., Pegram, R. G., Perry, B. D., Lemche, J., and Schels, H. F. (1988). The distribution of African swine fever virus isolated from Ornithodoros moubata in Zambia. *Epidemiol. Infect.* **101**(3), 547-64.

Woelk, C.H., Holmes, E.C. (2002). Reduced positive selection in vector-borne RNA viruses. *Mol. Biol. Evol.* **19**, 2333-2336.

Woese, C.R., Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088-5090.

Woolhouse, M.E., Webster, J.P., Domingo, E., Charlesworth, B., Levin, B.R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569-577.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97-159.

Xia, X. (1998). The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Mol. Biol. Evol.* **15**, 336-344.

Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y. (2003). An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26**, 1-7.

Xia, X., and Xie, Z. (2001). DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**(4), 371-3.

Yanez, R. J., and Vinuela, E. (1993). African swine fever virus encodes a DNA ligase. *Virology* **193**(1), 531-6.

Yanez, R. J., Boursnell, M., Nogal, M. L., Yuste, L., and Vinuela, E. (1993a). African swine fever virus encodes two genes which share significant homology with the two largest subunits of DNA-dependent RNA polymerases. *Nucleic Acids Res.* **21**(10), 2423-7.

Yanez, R. J., Rodriguez, J. M., Nogal, M. L., Yuste, L., Enriquez, C., Rodriguez, J. F., and Vinuela, E. (1995). Analysis of the complete nucleotide sequence of African swine fever virus. *Virology* **208**(1), 249-78.

Yanez, R. J., Rodriguez, J. M., Rodriguez, J. F., Salas, M. L., and Vinuela, E. (1993b). African swine fever virus thymidylate kinase gene: sequence and transcriptional mapping. *J. Gen. Virol.* **74** (Pt 8), 1633-8.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105-111.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306-314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367-372.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568-573.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.* **51**, 423-432.

Yang, Z. (2006). Computational molecular evolution. Oxford University Press, Oxford.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**(8), 1586-91.

Yoder, A. D., Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol*. **17**(7), 1081-1090.

Yourno, J., Conroy, J. (1992). A novel polymerase chain reaction method for detection of human immunodeficiency virus in dried blood spots on filter paper. *J. Clin. Microbiol*. **30**(11), 2887-2892.

Yozawa, T., Kutish, G. F., Afonso, C. L., Lu, Z., and Rock, D. L. (1994). Two novel multigene families, 530 and 300, in the terminal variable regions of African swine fever virus genome. *Virology* **202**(2), 997-1002.

Yutin, N., Koonin, E.V. (2009). Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. *Biol. Direct* **4**, 51.

Yutin, N., Wolf, Y.I., Raoult, D., Koonin, E.V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. J. Virol. **6**, 223.

Zhang, Z., Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338-5348.

Zhang, Z., Wang, Y., Wang, L., Gao, P. (2012). The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One* **5**, e14316.

Zheng, Q. (2001). On the dispersion index of a Markovian molecular clock. *Math. Biosci.* **172**, 115-128.

Zoller, S., Schneider, A. (2010). Empirical analysis of the most relevant parameters of codon substitution models. *J. Mol. Evol.* **70**, 605-612.

Zollner, B. (2004). Surveillance of the molecular epidemiology of hepatitis B virus in industrialized countries: necessary despite low prevalence and an available, effective vaccine? *Clin. Infect. Dis*. **39**(7), 953-954.

Zsak, L., Caler, E., Lu, Z., et al. (1998). A nonessential African swine fever virus gene UK is a significant virulence determinant in domestic swine. *J. Virol.* **72**, 1028-1035.

Zsak, L., Lu, Z., Burrage, T.G., et al. (2001). African swine fever virus multigene family 360 and 530 genes are novel macrophage host range determinants. *J. Virol.* **75**, 3066-3076.

Zsak, L., Onisk, D. V., Afonso, C. L., and Rock, D. L. (1993). Virulent African swine fever virus isolates are neutralized by swine immune serum and by monoclonal antibodies recognizing a 72-kDa viral protein. *Virology* **196**(2), 596-602.

Zuckerkandl, E., Pauling, L. (1962). Molecular disease, evolution, and genetic heterogeneity. In: Horyzons in Biochemestry, ed. M. Kasha and B. Pullman, pp. 189-225. New York: Academic Press.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**(2), 357-66.

ANNEXES

Annexe 1 : Isolats de virus PPA utilisés au cours de cette thèse. Les lignées sont indiquées selon notre nomenclature en lien avec les génotypes décrits dans la littérature.

Isolat	Pays	Date d'isolement	Espèce d'origine	B646L	E183L	CP204L	Génotype	Lignée	Publication
608	inconnu	NC	NC	-	-	AF462274	-	L1	Hernaez *et al.*, 2001
646	Espagne	1969	Porc domestique	FJ174351	FJ174392	-	I	L1.1	Gallardo *et al.*, 2009
1207	NC	NC	NC	-	-	AF462273	-	L1	Hernaez *et al.*, 2001
24823	Afrique du Sud	1975	NC	DQ250110	-	-	XX	L2.2.4	Boshoff *et al.*, 2007
04/Ol/02	Italie	2002	Porc domestique	-	FR681815	-	I	L1.1	Giammarioli *et al.*, 2011
1/Nu/97	Italie	1997	Porc domestique	FR668398	FR681812	-	I	L1.1	Giammarioli *et al.*, 2011
11/Og/04	Italie	2004	Porc domestique	FR668403	FR681817	-	I	L1.1	Giammarioli *et al.*, 2011
13/Nu/04	Italie	2004	Porc domestique	-	FR681818	-	I	L1.1	Giammarioli *et al.*, 2011
16/Og/04	Italie	2004	Porc domestique	FR682502	FR681819	-	I	L1.1	Giammarioli *et al.*, 2011
18/Nu/04	Italie	2004	Porc domestique	FR677326	FR681820	-	I	L1.1	Giammarioli *et al.*, 2011
2/Og/97	Italie	1997	Porc domestique	FR668399	FR681813	-	I	L1.1	Giammarioli *et al.*, 2011
22/Nu/04	Italie	2004	Porc domestique	FR668405	FR681821	-	I	L1.1	Giammarioli et al., 2011
23/Or/04	Italie	2004	Porc domestique	FR668406	FR681822	-	I	L1.1.5	Giammarioli *et al.*, 2011
24/Or/04	Italie	2004	Porc domestique	FR668407	FR681823	-	I	L1.1.5	Giammarioli et al., 2011
25/Nu/04	Italie	2004	Porc domestique	FR668408	FR681824	-	I	L1.1.5	Giammarioli *et al.*, 2011
26/Ss/04	Italie	2004	Porc domestique	-	FR681825	-	I	L1.1	Giammarioli *et al.*, 2011
3/Og/98	Italie	1998	Porc domestique	FR668400	FR681814	-	I	L1.1	Giammarioli *et al.*, 2011
30/Ol/04	Italie	2004	Porc domestique	FR668410	FR681826	-	I	L1.1.3	Giammarioli *et al.*, 2011
36/Ss/05	Italie	2005	Porc domestique	FR668411	FR681827	-	I	L1.1	Giammarioli *et al.*, 2011
38/Ss/07	Italie	2007	Porc domestique	FR668412	FR681828	-	I	L1.1	Giammarioli *et al.*, 2011
41/Og/07	Italie	2007	Sanglier	FR668413	FR681829	-	I	L1.1	Giammarioli *et al.*, 2011
42/Og/07	Italie	2007	Sanglier	FR668414	FR681830	-	I	L1.1	Giammarioli *et al.*, 2011
43/Og/07	Italie	2007	Sanglier	FR668415	FR681831	-	I	L1.1	Giammarioli *et al.*, 2011
46/Ca/08	Italie	2008	Porc domestique	-	FR681832	-	I	L1.1	Giammarioli *et al.*, 2011

47/Ss/08	Italie	2008	Porc domestique	-	FR681833	-	I	L1.1	Giammarioli *et al.*, 2011
48/Ss/08	Italie	2008	Porc domestique	-	FR681834	-	I	L1.1	Giammarioli *et al.*, 2011
5/Ca/02	Italie	2002	Porc domestique	FR668402	FR681816	-	I	L1.1.3	Giammarioli *et al.*, 2011
51/Nu/09	Italie	2009	Porc domestique	FR668419	FR681835	-	I	L1.1	Giammarioli *et al.*, 2011
52/Nu/09	Italie	2009	Porc domestique	-	FR681836	-	I	L1.1	Giammarioli *et al.*, 2011
53/Nu/09	Italie	2009	Porc domestique	FR677327	FR681837	-	I	L1.1	Giammarioli *et al.*, 2011
98/ASF/NG	Nigeria	1998	Porc domestique	AF159503	-	-	I	L1.1	Bastos *et al.*, 2003
Ali61	Espagne	1961	Porc domestique	FJ154445	FJ174384	-	I	L1.1	Gallardo *et al.*, 2009
Almodovar99	Portugal	1999	Porc domestique	DQ028306	DQ028315	-	I	L1.1	Duarte *et al.*, 2005
Almodovar99E2	Portugal	1999	Tique *Ornithodoros*	DQ028308	DQ028316	-	I	L1.1	Duarte *et al.*, 2005
Almodovar99NE1	Portugal	1999	Tique *Ornithodoros*	DQ028309	DQ028317	-	I	L1.1	Duarte *et al.*, 2005
Ambaton01	Madagascar	2001	Porc domestique	-	-	-	II	L1.2	Notre étude
Ambilo03	Madagascar	2003	Porc domestique	-	-	-	II	L1.2	Notre étude
Ambovo99	Madagascar	1999	Porc domestique	-	-	-	II	L1.2	Notre étude
Ampani99	Madagascar	1999	Porc domestique	-	-	-	II	L1.2	Notre étude
ANG/70	Angola	1970	Porc domestique	AF301542	EU874327	EU874271	I	L1.1	Bastos *et al.*, 2003
Ang72	Angola	1972	Porc domestique	FJ174378	FJ174424	-	I	L1.1	Gallardo *et al.*, 2009
Antana00	Madagascar	2000	Porc domestique	-	-	-	II	L1.2	Notre étude
Antani03	Madagascar	2003	Porc domestique	-	-	-	II	L1.2	Notre étude
Antsir99	Madagascar	1999	Porc domestique	-	-	-	II	L1.2	Notre étude
Antsira02	Madagascar	2002	Porc domestique	-	-	-	II	L1.2	Notre étude
Arivo01	Madagascar	2001	Porc domestique	-	-	-	II	L1.2	Notre étude
Av71	Espagne	1971	Porc domestique	FJ174349	FJ174391	-	I	L1.1	Gallardo *et al.*, 2009
Avara02	Madagascar	2002	Porc domestique	-	-	-	II	L1.2	Notre étude
Awoshie99	Ghana	1999	NC	AF504885	-	-	I	L1.1	Bastos *et al.*, 2003
B74	Espagne	1974	Porc domestique	FJ174350	FJ174393	-	I	L1.1	Gallardo *et al.*, 2009
Ba71V	Espagne	1971	NC	FJ174348	FJ174390	M96354	I	L1.1	Bastos *et al.*, 2003
BAN/91/1	Malawi	1991	*Sus scrofa*	AY351501	EU874348	EU874260	VIII	L3.1	Lubisi *et al.*, 2005
Barrancos93	Portugal	1993	Porc domestique	DQ028307	DQ028318	-	I	L1.1	duarte *et al.*, 2005

Bartlett2	Kenya	1959	Phaecochoerus aethiopicus	AY351532	-	-	X	L4.2.2.2.1	Lubisi *et al.*, 2005
BEL85	Belgique	1985	Porc domestique	AF449466	-	-	I	L1.1	Bastos *et al.*, 2003
BEN97/4	Benin	1997	Porc domestique	AY972164	-	-	I	L1.1	Phologane *et al.*, 2005
Benin97/1	Benin	1997	Porc domestique	AF302816	AM712239	AM712239	I	L1.1	Bastos *et al.*, 2003
Betrok99	Madagascar	1999	Porc domestique	-	-	-	II	L1.2	Notre étude
Bongera/83	Malawi	1983	Porc domestique	-	X84905	-	VIII	L3.1	Sun *et al.*, 1995
BOT/99/1	Botswana	1999	Porc domestique	-	EU874382	-	III	L2.2.1	Heath *et al.*, 2008
Brazil78	Brésil	1978	Porc domestique	FJ238537	FJ238535	-	I	L1.1	Gallardo *et al.*, 2009
Brazil79	Brésil	1979	NC	AF302809	-	-	I	L1.1	Bastos *et al.*, 2003
BUR/84/1	Burundi	1984	Porc domestique	AF449463	EU874364	EU874298	X	L4.2.2.2.3	Bastos *et al.*, 2003
BUR/84/2	Burundi	1984	Porc domestique	AF449464	-	-	X	L4.2.2.2.3	Bastos *et al.*, 2003
BUR/90/1	Burundi	1990	Porc domestique	AF449472	EU874363	EU874299	X	L4.2.2.2.3	Bastos *et al.*, 2003
BUR/90/3	Burundi	1990	*Sus scrofa*	AY351525	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
Ca04.1	Italie	2004	Porc domestique	FR668270	FR668271	-	I	L1.1	Giammarioli *et al.*, 2011
Ca78	Italie	1978	Porc domestique	FJ174357	FJ174401	-	I	L1.1	Nix *et al.*, 2006
Ca97	Italie	1997	Porc domestique	FJ174371	FJ174416	-	I	L1.1	Gallardo *et al.*, 2009
CAM/82	Cameroun	1982	Porc domestique	AF301544	-	-	I	L1.1	Bastos *et al.*, 2003
CAM/02/1	Cameroun	2002	NC	-	EU874323	EU874267	I	L1.1	Heath *et al.*, 2008
CAM/02/2	Cameroun	2002	NC	-	EU874324	EU874268	I	L1.1	Heath *et al.*, 2008
CAM/02/3	Cameroun	2002	NC	-	EU874325	EU874269	I	L1.1	Heath *et al.*, 2008
CAM/02/4	Cameroun	2002	NC	-	EU874326	EU874270	I	L1.1	Heath *et al.*, 2008
CAM/85/4	Cameroun	1985	NC	-	EU874322	EU874272	I	L1.1	Bastos *et al.*, 2003
CHG/88/1	Zambie	1988	*Sus scrofa*	AY351552	-	-	VIII	L3.1	Lubisi *et al.*, 2005
CHJ/89/1	Zambie	1989	*Sus scrofa*	AY351519	EU874346	EU874262	VIII	L3.1	Lubisi *et al.*, 2005
CHK/89/2	Zambie	1989	*Sus scrofa*	AY351526	-	-	VIII	L3.1	Lubisi *et al.*, 2005
CHM/88/1	Zambie	1988	*Sus scrofa*	AY351520	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Chrôme01	Madagascar	2001	Porc domestique	-	-	-	II	L1.2	Notre étude
CM96	Côte d'Ivoire	1996	Porc domestique	-	-	-	I	L1.1	Notre étude

Co61	Espagne	1961	Porc domestique	FJ174346	FJ174386	-	I	L1.1	Gallardo *et al.*, 2009
Co62	Espagne	1962	Porc domestique	FJ174347	FJ174387	-	I	L1.1	Gallardo *et al.*, 2009
Co68	Espagne	1968	Porc domestique	FJ238538	FJ174388	-	I	L1.1	Gallardo *et al.*, 2009
Coimbra87	Portugal	1987	Porc domestique	DQ028310	DQ028319	-	I	L1.1	Duarte *et al.*, 2005
Con09/Abo	RDC	2009	Porc domestique	-	HQ645951	-	IX	L4.1	Gallardo *et al.*, 2011
Con09/Bzz020	RDC	2009	Porc domestique	-	HQ645950	-	IX	L4.1	Gallardo *et al.*, 2011
Con09/Ni16	RDC	2009	Porc domestique	-	HQ645948	-	IX	L4.1	Gallardo *et al.*, 2011
Con09/Pk45	RDC	2009	Porc domestique	-	HQ645952	-	IX	L4.1	Gallardo *et al.*, 2011
Con09/PN003	RDC	2009	Porc domestique	-	HQ645949	-	IX	L4.1	Gallardo *et al.*, 2011
Cro5.3	Afrique du Sud	1996	Tique *Ornithodoros*	AY578691	-	-	NA	L2.3.4	Zsak *et al.*, 2005
Cro1.2	Afrique du Sud	1996	Tique *Ornithodoros*	AY578690	-	-	XX	L2.2.4	Zsak *et al.*, 2005
CV97	Cap Vert	1997	Porc domestique	FJ174380	FJ174427	-	I	L1.1	Gallardo *et al.*, 2009
CV98	Cap Vert	1998	Porc domestique	FJ174381	FJ174428	-	I	L1.1	Gallardo *et al.*, 2009
CVR/Tet20	Nigeria	2005	NC	GQ427180	-	-	I	L1.1	Owolodun *et al.*, 2010
CVR/Tet21	Nigeria	2005	NC	GQ427181	-	-	I	L1.1	Owolodun *et al.*, 2010
CVR/Tet29	Nigeria	2006	NC	GQ427183	-	-	I	L1.1	Owolodun *et al.*, 2010
Dakar59	Sénégal	1959		AF301538	-	-	I	L1.1	Bastos *et al.*, 2003
Davis	Kenya	1959	*Phaecochoerus aethiopicus*	AY351527	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
DED/89/1	Malawi	1989	*Sus scrofa*	AY351502	EU874349	EU874256	VIII	L3.1	Lubisi *et al.*, 2005
DED/91/1	Malawi	1991	*Sus scrofa*	AY351503	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Dedza	Malawi	1986	Porc domestique	AF449479	-	-	VIII	L3.1	Bastos *et al.*, 2003
Doig	Kenya	1957	*Phaecochoerus aethiopicus*	AY351528	-	-	X	L4.2.2.2.2	Lubisi *et al.*, 2005
DomRep79	République Dominicaine	1978	NC	AF302810	FJ238534	-	I	L1.1	Bastos *et al.*, 2003
DOWA	Malawi	1986	*Sus scrofa*	AY351509	EU874350	EU874261	VIII	L3.1	Lubisi *et al.*, 2005
DR2	République Dominicaine	1979	Porc domestique	ASVMCPC	-	-	I	L1.1	Yu *et al.*, 1996
DR78	République Dominicaine	1978	Porc domestique	-	-	-	I	L1.1	Notre étude

E70	Espagne	1970	Porc domestique	AY578692	FJ174389	AF462272	I	L1.1	Zsak *et al.*, 2005
E75	Espagne	1975	Porc domestique	AY578693	FJ174394	AF462271	I	L1.1	Zsak *et al.*, 2005
F6	Afrique du Sud	1996	Tique *Ornithodoros*	AY578694	-	-	XIX	L2.2.3	Zsak *et al.*, 2005
Faharet98	Madagascar	1998	Porc domestique	-	-	-	II	L1.2	Notre étude
Fandria01	Madagascar	2001	Porc domestique	-	-	-	II	L1.2	Notre étude
Fianara00	Madagascar	2000	Porc domestique	-	-	-	II	L1.2	Notre étude
Fr64	France	1964	Porc domestique	FJ174374	FJ174421	-	I	L1.1	Nix *et al*., 2006
GAM/1/00	Gambie	2000	Porc domestique	AF449478	-	-	I	L1.1	Bastos *et al.*, 2003
Gara08	Madagascar	2008	Porc domestique	-	-	-	II	L2.1	Notre étude
Gasson	Kenya	1961	*Sus scrofa*	AY351529	-	-	X	L4.2.2.2.1	Lubisi *et al.*, 2005
Georgia2007	Georgia	2007	Porc domestique	AM999764	AM999765	AM999766	II	L1.2	Rowland *et al.*, 2008
GHA/1/00	Ghana	2000	Porc domestique	AF504888	-	-	I	L1.1	Bastos *et al.*, 2003
Ghana	Ghana	2000	NC	AF504889	-	-	I	L1.1	Bastos *et al.*, 2003
GHANA/02/1	Ghana	2002	NC	-	EU874328	EU874293	I	L1.1	Heath *et al*., 2008
GR21/11	Afrique du Sud	1978	Tique *Ornithodoros*	FJ455836	-	-	XX	L2.2.4	Arnot *et al.*, 2009
GR21/23	Afrique du Sud	1978	Tique *Ornithodoros*	FJ455837	-	-	XX	L2.2.4	Arnot *et al.*, 2009
GR22/6	Afrique du Sud	1978	Tique *Ornithodoros*	FJ455838	-	-	XX	L2.2.4	Arnot *et al.*, 2009
GR44A2	Afrique du Sud	1979	Tique *Ornithodoros*	FJ455835	-	-	XX	L2.3.2	Arnot *et al.*, 2009
GUL/88/1	Zambie	1988	*Sus scrofa*	AY351521	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Hai81	Haïti	1981	Porc domestique	FJ174375	FJ238536	-	I	L1.1	Gallardo *et al.*, 2009
HindeI	Kenya	1954	Suidé	AY351530	-	-	X	L4.2.2.2.2	Lubisi *et al.*, 2005
HindeII	Kenya	1959	Porc domestique	AF449480	-	-	X	L4.2.2.2.2	Bastos *et al.*, 2003
HOL86	Pays.Bas	1986	Porc domestique	AF449467	-	-	I	L1.1	Bastos *et al.*, 2003
ht79	Haïti	1979	Porc domestique	AY578695	-	-	I	L1.1	Zsak *et al.*, 2005
Hu90	Espagne	1990	Porc domestique	FJ174355	FJ174399	-	I	L1.1	Gallardo *et al.*, 2009
Hu94	Espagne	1994	Porc domestique	FJ174356	FJ174400	-	I	L1.1	Gallardo *et al.*, 2009
IC/3/96	Côte d'Ivoire	1996	Porc domestique	AF504882	-	-	I	L1.1	Bastos *et al.*, 2003
IC/5/76	Côte d'Ivoire	1996	NC	AF504883	-	-	I	L1.1	Bastos *et al.*, 2003
IC/1/96	Côte d'Ivoire	1996	Porc domestique	AF302814	-	-	I	L1.1	Bastos *et al.*, 2003
IC/2/96	Côte d'Ivoire	1996	Porc domestique	AF302815	EU874319	EU874282	I	L1.1	Bastos *et al.*, 2003

IC96	Côte d'Ivoire	1996	Porc domestique	FJ174379	FJ174429	-	I	L1.1	Gallardo *et al.*, 2009
JON/89/13	Zambie	1989	Porc domestique	AF449469	-	-	VIII	L3.1	Bastos *et al.*, 2003
K1	Afrique du Sud	1996	Tique *Ornithodoros*	AY578696	-	-	III	L2.2.1	Zsak *et al.*, 2005
KAB/62	Zambie	1983	Tique *Ornithodoros*	AY351522	EU874331	EU874289	XI	L3.2	Lubisi *et al.*, 2005
KAB/94/1	Kenya	1994	Porc domestique	AY972163	-	-	X	L4.2.2.2.3	Phologane *et al.*, 2005
KAC/91/2	Malawi	1991	*Sus scrofa*	AY351504	-	-	VIII	L3.1	Lubisi *et al.*, 2005
KAL/88/1	Zambie	1988	Porc domestique	AF449468	-	-	VIII	L3.1	Bastos *et al.*, 2003
KANA/89/1	Zambie	1989	*Sus scrofa*	AY351523	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Kat67	RDC	1967	Porc domestique	FJ174377	FJ174423	-	I	L1.1	Gallardo *et al.*, 2009
Katanga63	RDC	1963	Porc domestique	AF301540	-	-	I	L1.1	Bastos *et al.*, 2003
KAV/89/1	Zambie	1989	Tique *Ornithodoros*	AF449470	-	-	VIII	L3.1	Bastos *et al.*, 2003
KEN/05/1	Kenya	2005	NC	-	EU874368	EU874301	IX	L4.1	Heath *et al.*, 2008
Ken05.DPk16	Kenya	2005	Porc domestique	HM745264	HM745347	HM745369	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPk18	Kenya	2005	Porc domestique	HM745265	HM745348	HM745370	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPk2	Kenya	2005	Porc domestique	HM745263	HM745346	HM745368	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPk21	Kenya	2005	Porc domestique	HM745266	HM745349	HM745371	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPk27	Kenya	2005	Porc domestique	HM745267	HM745350	HM745372	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPN15	Kenya	2005	Porc domestique	HM745269	HM745352	HM745374	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPN2	Kenya	2005	Porc domestique	HM745268	HM745351	HM745373	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPN23	Kenya	2005	Porc domestique	HM745270	HM745353	HM745375	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPU1	Kenya	2005	Porc domestique	HM745271	HM745354	HM745376	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPU11	Kenya	2005	Porc domestique	HM745273	HM745356	HM745378	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPU2	Kenya	2005	Porc domestique	HM745272	HM745355	HM745377	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05.DPU22	Kenya	2005	Porc domestique	HM745274	HM745357	HM745379	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05/Tk1	Kenya	2005	Tique *Ornithodoros*	HM745253	HM745336	HM745358	X	L4.2.1	Gallardo *et al.*, 2011
Ken05/Tk10	Kenya	2005	Tique *Ornithodoros*	HM745262	HM745345	HM745367	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05/Tk2	Kenya	2005	Tique *Ornithodoros*	HM745254	HM745337	HM745359	X	L4.2.1	Gallardo *et al.*, 2011
Ken05/Tk3	Kenya	2005	Tique *Ornithodoros*	HM745255	HM745338	HM745360	X	L4.2.1	Gallardo *et al.*, 2011
Ken05/Tk4	Kenya	2005	Tique *Ornithodoros*	HM745256	HM745339	HM745361	X	L4.2.1	Gallardo *et al.*, 2011

Ken05/Tk5	Kenya	2005	Tique *Ornithodoros*	HM745257	HM745340	HM745362	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05/Tk6	Kenya	2005	Tique *Ornithodoros*	HM745258	HM745341	HM745363	X	L4.2.1	Gallardo *et al.*, 2011
Ken05/Tk7	Kenya	2005	Tique *Ornithodoros*	HM745259	HM745342	HM745364	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken05/Tk8	Kenya	2005	Tique *Ornithodoros*	HM745260	HM745343	HM745365	X	L4.2.1	Gallardo *et al.*, 2011
Ken05/Tk9	Kenya	2005	Tique *Ornithodoros*	HM745261	HM745344	HM745366	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken06.B1	Kenya	2006	Porc domestique	FJ154434	FJ174441	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.B2	Kenya	2006	Porc domestique	FJ154435	FJ174442	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.B3	Kenya	2006	Porc domestique	FJ154436	FJ174443	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.B4	Kenya	2006	Porc domestique	FJ154437	FJ174444	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.B5	Kenya	2006	Porc domestique	FJ154438	FJ174445	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.Bus	Kenya	2006	Porc domestique	FJ154439	FJ174446	-	IX	L4.1	Gallardo *et al.*, 2009
Ken06.Kis	Kenya	2006	Porc domestique	FJ154440	FJ174447	-	IX	L4.1	Gallardo *et al.*, 2009
Ken07.Eld1	Kenya	2007	Porc domestique	FJ154441	FJ174438	-	IX	L4.1	Gallardo *et al.*, 2009
Ken07.Eld2	Kenya	2007	Porc domestique	FJ154442	FJ174439	-	IX	L4.1	Gallardo *et al.*, 2009
Ken07.Kia	Kenya	2007	Porc domestique	FJ154443	FJ174437	-	IX	L4.1	Gallardo *et al.*, 2009
Ken07.Nak	Kenya	2007	Porc domestique	FJ154444	FJ174440	-	IX	L4.1	Gallardo *et al.*, 2009
Ken08Tk.2/1	Kenya	2008	Tique *Ornithodoros*	HM745275	HM745323	HM745380	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken08Tk.2/3	Kenya	2008	Tique *Ornithodoros*	HM745276	HM745324	HM745381	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken08WH/4	Kenya	2008	Phacochoerus africanus	HM745285	HM745333	HM745390	IX	L4.1	Gallardo *et al.*, 2011
Ken08WH/5	Kenya	2008	Phacochoerus africanus	HM745286	HM745334	HM745392	IX	L4.1	Gallardo *et al.*, 2011
Ken08WH/8	Kenya	2008	Phacochoerus africanus	HM745287	HM745335	HM745391	IX	L4.1	Gallardo *et al.*, 2011
Ken09Tk.13/1	Kenya	2009	Tique *Ornithodoros*	HM745277	HM745325	HM745382	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.13/2	Kenya	2009	Tique *Ornithodoros*	HM745278	HM745326	HM745383	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.15/4	Kenya	2009	Tique *Ornithodoros*	HM745279	HM745327	HM745384	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.15/6	Kenya	2009	Tique *Ornithodoros*	HM745280	HM745328	HM745385	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.19/11	Kenya	2009	Tique *Ornithodoros*	HM745283	HM745331	HM745388	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.19/2	Kenya	2009	Tique *Ornithodoros*	HM745281	HM745329	HM745386	X	L4.2.2.1	Gallardo *et al.*, 2011
Ken09Tk.19/7	Kenya	2009	Tique *Ornithodoros*	HM745282	HM745330	HM745387	X	L4.2.2.1	Gallardo *et al.*, 2011

Ken09Tk.20/5	Kenya	2009	Tique *Ornithodoros*	HM745284	HM745332	HM745389	X	L4.2.2.1	Gallardo *et al.*, 2011
Kenya1950	Kenya	1950	Porc domestique	AY261360	EU874353	EU874297	X	L4.2.2.2.3	Kutish *et al.*, 2003
ker64	Kenya	1964	Porc domestique	AY578697	-	-	I	L1.1	Zsak *et al.*, 2005
KilleanI	Kenya	1959	*Phaecochoerus aethiopicus*	AY351550	-	-	X	L4.2.2.2.2	Lubisi *et al.*, 2005
KilleanII	Kenya	1959	*Phaecochoerus aethiopicus*	AY351551	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
KilleanIII	Kenya	1959	*Phaecochoerus aethiopicus*	AY351531	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
KimakiaI	Kenya	1961	*Phaecochoerus porcus*	AY351533	-	-	I	L1.1	Lubisi *et al.*, 2005
KimakiaII	Kenya	1961	*Phaecochoerus porcus*	AY351534	-	-	I	L1.1	Lubisi *et al.*, 2005
KIRT/89/2	Tanzanie	1989	Tique *Ornithodoros*	AY351511	-	-	I	L3.1	Lubisi *et al.*, 2005
KIRT/89/3	Tanzanie	1989	Tique *Ornithodoros*	AY351512	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
KIRT/89/4	Tanzanie	1989	Tique *Ornithodoros*	AY351513	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
KIRW/89/1	Tanzanie	1989	*Phaecochoerus aethiopicus*	AY351514	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
KLI/88/2	Zambie	1988	*Sus scrofa*	AY351553	EU874347	EU874258	VIII	L3.1	Lubisi *et al.*, 2005
kn66	Kenya	1966	Porc domestique	AY578698	-	-	X	L4.2.2.2.3	Zsak *et al.*, 2005
Kwh/12	Tanzanie	1968	*Phaecochoerus*	AF301546	-	-	X	L4.2.2.2.3	Bastos *et al.*, 2003
LIL/89/1	Malawi	1989	*Sus scrofa*	AY351505	-	-	VIII	L3.1	Lubisi *et al.*, 2005
LIL/90/1	Malawi	1990	*Sus scrofa*	AY351510	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Lillie	Afrique du Sud	1973	Tique *Ornithodoros*	DQ250109	EU874341	EU874306	XX	L2.2.4	Boshoff *et al.*, 2007
Lis57	Portugal	1957	Porc domestique	AF301537	FJ174420	-	I	L1.1	Bastos *et al.*, 2003
Lis60	Portugal	1960	Porc domestique	AF301539	X84889	EU874273	I	L1.1	Bastos *et al.*, 2003

LIV/10/11	Zambie	1983	Tique *Ornithodoros*	AY351535	-	-	I	L1.1.1	Lubisi *et al.*, 2005
LIV/12/17	Zambie	1983	Tique *Ornithodoros*	AY351524	-	-	I	L1.1.2	Lubisi *et al.*, 2005
LIV/13/33	Zambie	1983	Tique *Ornithodoros*	AY494560	-	-	I	L1.1	Lubisi *et al.*, 2005
LIV/5/4	Zambie	1983	Tique *Ornithodoros*	AY351537	-	-	I	L1.1.2	Lubisi *et al.*, 2005
LIV/9/31	Zambie	1983	Tique *Ornithodoros*	AY351538	-	-	I	L1.1.2	Lubisi *et al.*, 2005
LIV/9/35	Zambie	1983	Tique *Ornithodoros*	AY351539	-	-	I	L1.1.2	Lubisi *et al.*, 2005
LIV5/40	Zambie	1982	Tique *Ornithodoros*	AY351536	-	-	I	L1.1.2	Lubisi *et al.*, 2005
LUS/93/1	Zambie	1991	*Sus scrofa*	AY351563	EU874377	EU874275	I	L1.2	Lubisi *et al.*, 2005
M1	Afrique du Sud	1966	Tique *Ornithodoros*	AY578699	-	-	XIX	L2.2.3	Zsak *et al*., 2005
M61	Espagne	1961	Porc domestique	FJ174345	FJ174385	-	I	L1.1	gallardo *et al.*, 2009
MAD/1/1998	Madagascar	1998	Porc domestique	AF270706	-	-	II	L1.2	Bastos *et al.*, 2003
Madrid/62	Espagne	1962	NC	AF449461	-	-	I	L1.1	Bastos *et al.*, 2003
Mafra86	Portugal	1986	Porc domestique	DQ028312	DQ028321	-	I	L1.1	Duarte *et al.*, 2005
Mahaja02	Madagascar	2002	Porc domestique	-	-	-	II	L1.2	Notre étude
MAL/2002/1	Malawi	2002	*Sus scrofa*	AY494553	EU874373	EU874311	V	L2.1.1	Lubisi *et al.*, 2005
MAL/1978	Malawi	1978		AF270707	-	-	VIII	L3.1	Bastos *et al.*, 2003
Malte78	Malte	1978	Porc domestique	AF301543	FJ174419	-	I	L1.1	Bastos *et al.*, 2003
MAN/89/2	Zambie	1989	*Sus scrofa*	AY351562	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Marovo02	Madagascar	2002	Porc domestique	-	-	-	II	L1.2	Notre étude
MAU/2007/1	Maurice	2007	Porc domestique	FJ528594	-	-	I	L1.2	Lubisi *et al.*, 2009
MAU/2008/1	Maurice	2008	Porc domestique	FJ528595	-	-	I	L1.2	Lubisi *et al.*, 2009
MCH/89/1	Malawi	1989	*Sus scrofa*	AY351506	EU874352	-	VIII	L3.1	Lubisi *et al.*, 2005

MCH/89/3	Malawi	1989	*Sus scrofa*	AY351507	EU874351	EU874259	VIII	L3.1	Lubisi *et al.*, 2005
Mchinji075	Malawi	1997	*Sus scrofa*	AY351508	-	-	VIII	L3.1	Lubisi *et al.*, 2005
MFUE/6/1	Zambie	1982	Tique *Ornithodoros*	AY351561	-	-	XII	L3.3	Lubisi *et al.*, 2005
MHC/89/1	Malawi	1989	Porc domestique	-	-	EU874292	VIII	L3.1	Heath *et al.*, 2008
Mkuzi78	Afrique du Sud	1978	Tique *Ornithodoros*	AY578700	-	-	I	L1.1.1	Zsak *et al.*, 2005
MKUZI79	Afrique du Sud	1979	Tique *Ornithodoros*	AY261362	EU874367	EU874294	I	L1.1.1	Kutish *et al.*, 2003
Morama98	Madagascar	1998	Porc domestique	-	-	-	II	L1.2	Notre étude
Moronda02	Madagascar	2002	Porc domestique	-	-	-	II	L1.2	Notre étude
MOZ/02/1	Mozambique	2002	Porc domestique	-	EU874380	EU874315	II	L1.2	Heath *et al.*, 2008
MOZ/02/2	Mozambique	2002	Porc domestique	-	EU874376	EU874274	II	L1.2	Heath *et al.*, 2008
MOZ/03/1	Mozambique	2003	NC	-	-	EU874314	II	L1.2	Heath *et al.*, 2008
MOZ/05/1	Mozambique	2005	NC	-	-	EU874313	V/VI	L2.1	Heath *et al.*, 2008
Moz/1/03	Mozambique	2003	NC	FJ175199	EU874379	-	II	L1.2	Heath *et al.*, 2008
MOZ/94/1	Mozambique	1994	Porc domestique	AF270711	-	-	VI	L2.1.2	Bastos *et al.*, 2003
Moz/1/05	Mozambique	2005	NC	FJ175200	-	-	II	L1.2	Heath *et al.*, 2008
MOZ/1/98	Mozambique	1998	NC	AF270705	-	-	VIII	L3.1	Bastos *et al.*, 2003
MOZ/1960	Mozambique	1960	Porc domestique	AF270708	EU874371	EU874309	V	L2.1.1	Bastos *et al.*, 2004
MOZ/1979	Mozambique	1979	Porc domestique	AF270709	EU874372	EU874310	V	L2.1.1	Bastos *et al.*, 2004
MOZ/94/1	Mozambique	1994	Porc domestique	-	EU874342	EU874263	V/VI	L2.1	Heath *et al.*, 2008
MOZ/94/8	Mozambique	1994	Porc domestique	-	EU874343	EU874276	V/VI	L2.1	Bastos *et al.*, 2003
Moz/98/1	Mozambique	1998	NC	-	EU874385	EU874317	VIII	L3.1	Heath *et al.*, 2008
Moz/1/05	Mozambique	2005	NC	-	EU874378	-	-	L1	Heath *et al.*, 2008
Moz64	Mozambique	1964	Porc domestique	FJ174376	FJ174422	-	V	L2.1.1	Gallardo *et al.*, 2009
MOZ/2001/1	Mozambique	2001	*Sus scrofa*	AY351516	-	-	VIII	L3.1	Lubisi *et al.*, 2005
MOZ/2002/1	Mozambique	2002	*Sus scrofa*	AY351517	-	-	II	L1.2	Lubisi *et al.*, 2005
MOZ/2002/2	Mozambique	2002	*Sus scrofa*	AY351518	-	-	II	L1.2	Lubisi *et al.*, 2005
MOZ/60/98	Mozambique	1998	*Sus scrofa*	AY274455	-	-	II	L1.2	Bastos *et al.*, 2004
MOZ/61/98	Mozambique	1998	*Sus scrofa*	AY274456	-	-	II	L1.2	Bastos *et al.*, 2004
MOZ/62/98	Mozambique	1998	*Sus scrofa*	AY274457	-	-	VIII	L3.1	Bastos *et al.*, 2004

MOZ/63/98	Mozambique	1998	*Sus scrofa*	AY274458	-	-	II	L1.2	Bastos *et al.*, 2004
MOZ/70/98	Mozambique	1998	*Sus scrofa*	AY274459	-	-	II	L1.2	Bastos *et al.*, 2004
MOZ/77/98	Mozambique	1998	*Sus scrofa*	AY538726	-	-	II	L1.2	Bastos *et al.*, 2004
MOZ/94/8	Mozambique	1994	Porc domestique	AF270712	-	-	VI	L2.1.2	Bastos *et al.*, 2004
MOZ/A/98	Mozambique	1998	*Sus scrofa*	AY274452	-	-	VIII	L3.1	Bastos *et al.*, 2004
MOZ/B/98	Mozambique	1998	*Sus scrofa*	AY274453	-	-	VIII	L3.1	Bastos *et al.*, 2004
MOZ/C/98	Mozambique	1998	*Sus scrofa*	AY274454	-	-	VIII	L3.1	Bastos *et al.*, 2004
MPI/89/1	Zambie	1989	*Sus scrofa*	AY351540	-	-	VIII	L3.1	Lubisi *et al.*, 2005
MPO89/1	Zambie	1989	*Sus scrofa*	AY351541	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Mu82	Espagne	1982	Porc domestique	FJ174352	FJ174395	-	I	L1.1	Gallardo *et al.*, 2009
MUR/07/1	NC	2007	NC	-	EU874384	EU874316	II	L1.2	Heath *et al.*, 2008
MWHOG/1	Kenya	1959	*Phaecochoerus aethiopicus*	AY351548	-	-	X	L4.2.2.2.3	Lubisi *et al.*, 2005
MWHOG/3	Kenya	1959	NC	AY351549	-	-	X	L4.2.1	Lubisi *et al.*, 2005
MWHOG/9	Kenya	1959	*Phaecochoerus aethiopicus*	AY351565	-	-	X	L4.2.2.1	Lubisi *et al.*, 2005
MwLIL20/1	Malawi	1983	Tique *Ornithodoros*	L00966	FJ174425	-	VIII	L3.1	Gallardo *et al.*, 2009
MZI/92/1	Malawi	1992	*Sus scrofa*	AY351543	-	EU874288	XII	L3.3	Lubisi *et al.*, 2005
MZI/94/1	Malawi	1994	NC	-	EU874360	-	-	L3	Heath *et al.*, 2008
NAM/1/80	Namibie	1980	Phacochoerus	AF504881	-	-	I	L1.1	Bastos *et al.*, 2003
NAM/1/95	Namibie	1995	NC	DQ250122	-	-	XVIII	L1.4	Boshoff *et al.*, 2007
NDA/1/90	Malawi	1990	Porc domestique	AF449473	-	-	VIII	L3.1	Bastos *et al.*, 2003
NGE/92/1	Malawi	1992	*Sus scrofa*	AY351544	-	-	VIII	L3.1	Lubisi *et al.*, 2005
NH/P68	Portugal	1968	Porc domestique	DQ028313	DQ028322	-	I	L1.1	Duarte *et al.*, 2005
NIG/6	Nigeria	1998	Porc domestique	AF270714	-	-	I	L1.1	Bastos *et al.*, 2003
NIG/1/99	Nigeria	1999	Porc domestique	AF504887	-	-	I	L1.1	Bastos *et al.*, 2003
NIG/2/98	Nigeria	1998	Porc domestique	AY972161	-	-	I	L1.1	Phologane *et al.*, 2005
NIG/3/98	Nigeria	1998	Porc domestique	AY972162	-	-	I	L1.1	Phologane *et al.*, 2005
NIG/01/1	Nigeria	2001	Porc domestique	-	EU874320	EU874277	I	L1.1	Heath *et al.*, 2008
Nig01	Nigeria	2001	Porc domestique	FJ174382	FJ174426	-	I	L1.1	Gallardo *et al.*, 2009

NKZ88/1	Zambie	1988	*Sus scrofa*	AY351554	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Nu04.3	Italie	2004	Porc domestique	FR668262	FR668247	-	I	L1.1	Giammarioli *et al.*, 2011
Nu04.4	Italie	2004	Porc domestique	FR668263	FR668248	-	I	L1.1	Giammarioli *et al.*, 2011
Nu04.6a	Italie	2004	Porc domestique	FR668264	FR668249	-	I	L1.1	Giammarioli *et al.*, 2011
Nu04.6b	Italie	2004	Porc domestique	FR668265	FR668250	-	I	L1.1	Giammarioli *et al.*, 2011
Nu04WB	Italie	2004	Sanglier	FR668269	FR668254	-	I	L1.1	Giammarioli *et al.*, 2011
Nu81	Italie	1981	Porc domestique	FJ174358	FJ174402	-	I	L1.1	Gallardo *et al.*, 2009
Nu90.1	Italie	1990	Porc domestique	AF302813	FJ174408	-	I	L1.1	Nix *et al.*, 2006
Nu91.3	Italie	1991	Porc domestique	FJ174364	FJ174409	-	I	L1.1	Gallardo *et al.*, 2009
Nu91.5	Italie	1991	NC	FJ174365	FJ174410	-	I	L1.1	Gallardo *et al.*, 2009
Nu93	Italie	1993	Porc domestique	FJ174366	FJ174411	-	I	L1.1	Gallardo *et al.*, 2009
Nu95.1	Italie	1995	Porc domestique	FJ174368	FJ174413	-	I	L1.1	Gallardo *et al.*, 2009
Nu96	Italie	1996	Porc domestique	FJ174369	FJ174414	-	I	L1.1	Gallardo *et al.*, 2009
Nu97	Italie	1997	Porc domestique	FJ174370	FJ174415	-	I	L1.1	Gallardo *et al.*, 2009
Nu98.3	Italie	1998	Porc domestique	FJ174372	FJ174417	-	I	L1.1	Gallardo *et al.*, 2009
Nu98.8B	Italie	1998	Porc domestique	FJ174373	FJ174418	-	I	L1.1	Gallardo *et al.*, 2009
NUR/90/1	Italie	1990	NC	AF302813	-	-	I	L1.1	Bastos *et al.*, 2003
NYA/1/2	Zambie	1986	Tique *Ornithodoros*	AY351555	EU874330	EU874302	XIV	-	Lubisi *et al.*, 2005
o1	Afrique du Sud	1996	Tique *Ornithodoros*	AY578701	-	-	XIX	L2.2.3	Zsak *et al.*, 2005
Ori84	Italie	1984	Porc domestique	FJ174360	FJ174404	-	I	L1.1	Nix *et al.*, 2006
Ori85	Italie	1985	Porc domestique	FJ174361	FJ174405	-	I	L1.1	Nix *et al.*, 2006
Ori90	Italie	1990	Porc domestique	FJ174363	FJ174407	-	I	L1.1	Nix *et al.*, 2006
Ori93	Italie	1993	Porc domestique	FJ174367	FJ174412	-	I	L1.1	Gallardo *et al.*, 2009
OURT88/1	Portugal	1988	Tique *Ornithodoros*	AF302811	-	-	I	L1.1	Bastos *et al.*, 2003
OURT88/3	Portugal	1988	Tique *Ornithodoros*	AM712240	AM712240	AM712240	I	L1.1	Nix *et al.*, 2006
PHW/88/1	Zambie	1988	*Sus scrofa*	AY351567	EU874366	EU874257	VIII	L3.1	Lubisi *et al.*, 2005
Portalegre90	Portugal	1990	Porc domestique	DQ028314	DQ028323	-	I	L1.1	Duarte *et al*., 2005
Pr4	Afrique du Sud	1996	Tique *Ornithodoros*	-	AY261363	AY261363	XX	L2.2.4	Kutish *et al*., 2003
RSA/03/1	Afrique du Sud	2003	NC	-	EU874333	EU874278	XX	L2.2.4	Heath *et al*., 2008

RSA/03/2	Afrique du Sud	2003	NC	-	EU874334	EU874279	III	L2.2.1	Heath *et al.*, 2008
RSA/03/3	Afrique du Sud	2003	NC	-	EU874337	EU874280	III	L2.2.1	Heath *et al.*, 2008
RSA/04/1	Afrique du Sud	2004	NC	-	EU874336	EU874284	III	L2.2.1	Heath *et al.*, 2008
RSA/04/3	Afrique du Sud	2004	NC	-	EU874370	EU874308	IV	L2.2.2	Heath *et al.*, 2008
RSA/07/1	Afrique du Sud	2007	NC	-	EU874383	-	-	L2	Heath *et al.*, 2008
RSA/1/98	Afrique du Sud	1998	Porc domestique	AF302818	-	-	VII	L2.3.1	Bastos *et al.*, 2003
RSA/1/99W	Afrique du Sud	1999	Phacochoerus	AF449477	EU874369	EU874307	IV	L2.2.2	Bastos *et al.*, 2003
RSA/95/1	Afrique du Sud	1995	NC	DQ250123	EU874340	EU874266	XX	L2.2.4	Boshoff *et al.*, 2007
RSA/95/4	Afrique du Sud	1995	NC	-	EU874332	EU874295	XX	L2.2.4	Heath *et al.*, 2008
RSA/95/5	Afrique du Sud	1995	NC	DQ250124	EU874358	EU874286	III	L2.2.1	Boshoff *et al.*, 2007
RSA/96/1	Afrique du Sud	1996	NC	DQ250125	EU874339	-	XXI	L2.3.3	Boshoff *et al.*, 2007
RSA/96/2	Afrique du Sud	1996	NC	DQ250126	EU874335	EU874281	XIX	L2.2.3	Boshoff *et al.*, 2007
RSA/96/3	Afrique du Sud	1996	NC	DQ250127	EU874375	EU874283	XIX	L2.2.3	Boshoff *et al.*, 2007
RSA/98/1	Afrique du Sud	1998	Porc domestique	-	EU874374	EU874312	VII	L2.3.1	Heath *et al.*, 2008
Sa88	Espagne	1988	Porc domestique	FJ174353	FJ174398	-	I	L1.1	Gallardo *et al.*, 2009
SAL/92/1	Malawi	1992	*Sus scrofa*	AY351546	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Se88	Espagne	1988	Porc domestique	FJ174354	FJ174397	-	I	L1.1	Gallardo *et al.*, 2009
SIY/91/2	Malawi	1991	*Sus scrofa*	AY351566	-	-	VIII	L3.1	Lubisi *et al.*, 2005
SPEC/120	Afrique du Sud	1987	NC	AF302812	-	-	XIX	L2.2.3	Bastos *et al.*, 2000
SPEC/125	Afrique du Sud	1987	NC	DQ250112	-	-	XIX	L2.2.3	Boshoff *et al.*, 2007
SPEC/140	Afrique du Sud	1987	Tique *Ornithodoros*	FJ455840	-	-	III	L2.2.1	Arnot *et al.*, 2009
SPEC/154	Botswana	1987	NC	DQ250113	EU874359	EU874291	VII	L2.3.1	Boshoff *et al.*, 2007
SPEC/205	Namibie	1989	NC	DQ250114	EU874329	EU874305	I	L1.1.1	Boshoff *et al.*, 2007
SPEC/207	Namibie	1989	NC	DQ250115	-	-	I	L1.1.1	Boshoff *et al.*, 2007
SPEC/209	Namibie	1989	NC	DQ250116	EU874365	EU874290	I	L1.1.1	Boshoff *et al.*, 2007
SPEC/245	Afrique du Sud	1992	NC	DQ250117	EU874381	-	XXII	L2.3.2	Boshoff *et al.*, 2007
SPEC/251	Afrique du Sud	1992	NC	DQ250118	-	-	XIX	L2.2.3	Boshoff *et al.*, 2007
SPEC/257	Afrique du Sud	1993	NC	DQ250120	EU874338	EU874265	III	L2.2.1	Boshoff *et al.*, 2007
SPEC/260	Afrique du Sud	1993	NC	DQ250121	-	-	VII	L2.3.1	Boshoff *et al.*, 2007
SPEC/265	Mozambique	1994	Porc domestique	AF270710	EU874344	EU874264	VI	L2.1.2	Bastos *et al.*, 2003

SPEC/53	Afrique du Sud	1983	NC	DQ250111	-	-	XXI	L2.3.3	Boshoff *et al.*, 2007
SPEC/57	Afrique du Sud	1985	Tique *Ornithodoros*	FJ455839	-	-	III	L2.2.1	Arnot *et al.*, 2009
Ss04.10	Italie	2004	Porc domestique	FR668266	FR668251	-	I	L1.1	Giammarioli *et al.*, 2011
Ss05.3a	Italie	2005	Porc domestique	FR668267	FR668252	-	I	L1.1	Giammarioli *et al.*, 2011
Ss05.3b	Italie	2005	Porc domestique	FR668268	FR668253	-	I	L1.1	Giammarioli *et al.*, 2011
SS81	Italie	1981	Porc domestique	FJ174359	FJ174403	-	I	L1.1	Nix *et al.*, 2006
Ss88	Italie	1988	Porc domestique	FJ174362	FJ174406	-	I	L1.1	Nix *et al.*, 2006
SUM/14/11	Zambie	1983	Tique *Ornithodoros*	AY351542	EU874357	EU874287	XIII	L3.4	Lubisi *et al.*, 2005
TAN/01/1	Tanzanie	2001	*Sus scrofa*	AY494552	EU874356	EU874303	XV	L3.5	Lubisi *et al.*, 2005
TAN/02/3	Tanzanie	2002	NC	-	EU874355	-	XVI	L3.6	Heath *et al.*, 2008
TAN/03/1	Tanzanie	2003	*Sus scrofa*	AY494550	EU874354	EU874304	XVI	L3.6	Lubisi *et al.*, 2005
TAN/08/MABIBO	Tanzanie	2008	Porc domestique	-	GQ410768	-	-	L3.7	Mazinso *et al.*, 2011
TAN/08/MAZIMBU	Tanzanie	2008	Porc domestique	GQ410765	GO410767	-	XV	L3.7	Mazinso *et al.*, 2011
TAN/08/TURIANI	Tanzanie	2008	Porc domestique	-	GQ410771	-	-	L3.7	Mazinso *et al.*, 2011
TAN/03/2	Tanzanie	2003	*Sus scrofa*	AY494551	-	EU874255	XVI	L3.6	Lubisi *et al.*, 2005
Tanzania/87	Tanzanie	1987	Porc domestique	-	X84891	-	XV	L3.5	Sun *et al.*, 1995
Tengani62	Malawi	1962	Porc domestique	-	EU874318	EU874296	V/VI	L2.1	Heath *et al.*, 2008
TEN/89/1	Zambie	1989	*Sus scrofa*	AY351556	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Tengani60	Malawi	1960	Phacochoerus	AF301541	-	-	V	L2.1.1	Bastos *et al.*, 2003
THY/90/1	Malawi	1990	*Sus scrofa*	AY351545	-	-	VIII	L3.1	Lubisi *et al.*, 2005
TMB/89/1	Zambie	1989	*Sus scrofa*	AY351557	EU874361	EU874285	VIII	L3.1	Lubisi *et al.*, 2005
Togo/98	Togo	1998	NC	AF449481	-	-	I	L1.1	Bastos *et al.*, 2003
Tolagna01	Madagascar	2001	Porc domestique	-	-	-	II	L1.2	Notre étude
Tolagna99	Madagascar	1999	Porc domestique	-	-	-	II	L1.2	Notre étude
Toliar98	Madagascar	1998	Porc domestique	-	-	-	II	L1.2	Notre étude
Trench	Kenya	1959	*Phaecochoerus aethiopicus*	AY351547	-	-	X	L4.2.2.2.2	Lubisi *et al.*, 2005
Tsididy08	Madagascar	2008	Porc domestique	-	-	-	II	L1.2	Notre étude
Ug03H.1	Uganda	2003	Porc domestique	FJ154428	FJ174431	-	IX	L4.1	Gallardo *et al.*, 2009

Ug03H.2	Uganda	2003	Porc domestique	FJ154429	FJ174432	-	IX	L4.1	Gallardo *et al.*, 2009
Ug03H.3	Uganda	2003	Porc domestique	FJ154430	FJ174433	-	IX	L4.1	Gallardo *et al.*, 2009
Ug03P.4	Uganda	2003	Porc domestique	FJ154431	FJ174434	-	IX	L4.1	Gallardo *et al.*, 2009
Ug03P.5	Uganda	2003	Porc domestique	FJ154432	FJ174435	-	IX	L4.1	Gallardo *et al.*, 2009
Ug03P.6	Uganda	2003	Porc domestique	FJ154433	FJ174436	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.F7	Uganda	2007	Porc domestique	GQ477143	GQ477150	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.F8	Uganda	2007	Porc domestique	GQ477144	GQ477151	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.Mukono	Uganda	2007	Porc domestique	GQ477142	GQ477149	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.Wak1	Uganda	2007	Porc domestique	GQ477138	GQ477145	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.Wak2	Uganda	2007	Porc domestique	GQ477139	GQ477146	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.Wak3	Uganda	2007	Porc domestique	GQ477140	GQ477147	-	IX	L4.1	Gallardo *et al.*, 2009
UG07.Wak4	Uganda	2007	Porc domestique	GQ477141	GQ477148	-	IX	L4.1	Gallardo *et al.*, 2009
Ug64	Uganda	1964	Porc domestique	FJ174383	FJ174430	-	X	L4.2.2.2.2	Gallardo *et al.*, 2009
UGA/95/1	Uganda	1995	Porc domestique	-	EU874362	EU874300	IX	L4.1	Heath *et al.*, 2008
UGA/95/3	Uganda	1995	Porc domestique	AF449476	-	-	X	L4.2.2.2.3	Bastos *et al.*, 2003
UGA2003/1	Uganda	2003	*Sus scrofa*	AY351564	-	-	IX	L4.1	Lubisi *et al.*, 2005
Uganda	Uganda	1965	Porc domestique	L27499	-	-	X	L4.2.2.2.2	Bastos *et al.*, 2003
UgH03	Uganda	2003	Porc domestique	EF121429	-	-	IX	L4.1	Blanco *et al.*, 2006
UgP03	Uganda	2003	Porc domestique	-	-	-	IX	L4.1	Notre étude
Val76	Espagne	1976	Porc domestique	AF449462	-	-	I	L1.1	Bastos *et al.*, 2003
vic	Zimbabwe	1983	Porc domestique	AY578705	-	-	I	L1.1	Zsak *et al.*, 2005
VICT/90/1	Zimbabwe	1990	Tique *Ornithodoros*	AF449474	-	-	I	L1.1.1	Bastos *et al.*, 2003
Warmbaths	Afrique du Sud	1987	Tique *Ornithodoros*	AY261365	AY261365	AY261365	III	L2.2.1	Kutish *et al.*, 2003
wart	Namibie	1980	Phacochoerus	AY578706	-	-	IV	L2.2.2	Zsak *et al.*, 2005
Warthog	Namibie	1980	Phacochoerus	AY261366	AY261366	AY261366	IV	L2.2.2	Kutish *et al.*, 2003
wb	Afrique du Sud	1987	Tique *Ornithodoros*	AY578707	-	-	III	L2.2.1	Zsak *et al.*, 2005
YEL/88/4	Zambie	1988	*Sus scrofa*	AY351558	-	-	VIII	L3.1	Lubisi *et al.*, 2005
Z85	Espagne	1985	Porc domestique	AF449465	FJ174396	-	I	L1.1	Bastos *et al.*, 2003
Za	RDC	1987	Porc domestique	AY578708	-	-	XIX	L2.2.3	Zsak *et al.*, 2005
ZAM/01/1	Zambie	2001	*Sus scrofa*	AY494554	-	-	I	L1.1.4	Lubisi *et al.*, 2005

ZAM/02/1	Zambie	2002	*Sus scrofa*	AY494559	-	-	I	L1.1.4	Lubisi *et al.*, 2005
ZAM/01/2	Zambie	2001	*Sus scrofa*	AY494555	-	-	I	L1.1.4	Lubisi *et al.*, 2005
ZAM/01/3	Zambie	2001	*Sus scrofa*	AY494556	-	-	I	L1.1.4	Lubisi *et al.*, 2005
ZAM/01/4	Zambie	2001	*Sus scrofa*	AY494557	-	-	I	L1.1.4	Lubisi *et al.*, 2005
ZAM/01/5	Zambie	2001	*Sus scrofa*	AY494558	-	-	I	L1.1.4	Lubisi *et al.*, 2005
ZAM/88/1	Zambie	1988	*Sus scrofa*	AY351559	-	-	VIII	L3.1	Lubisi *et al.*, 2005
ZIM/92/1	Zimbabwe	1992	NC	DQ250119	EU874345	-	XII	L1.3	Boshoff *et al.*, 2007
ZON/88/1	Zambie	1988	*Sus scrofa*	AY351560	-	-	VIII	L3.1	Lubisi *et al.*, 2005

Annexe 2 : Fichier de contrôle du logiciel PAML pour l'analyse d_N/d_S des trois gènes du virus PPA étudiés.

```
noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
verbose = 1  * 0: concise ; 1: detailed, 2: too much
runmode = 0  * 0: user tree ;  1: semi-automatic ;  2: automatic
* 3: StepwiseAddition ; (4,5):PerturbationNNI ; -2: pairwise

seqtype = 1  * 1:codons ; 2:AAs ; 3:codons-->AAs
CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

*       ndata = 1
aaDist = 0  * 0:equal, +:geometric ; -:linear, 1-6:G1974,Miyata,c,p,v,a
aaRatefile = dat/jones.dat  * only used for aa seqs with model=empirical(_F)
             * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

model = 0
            * models for codons:
            * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
            * models for AAs or codon-translated AAs:
            * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
            * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

NSsites = 8  * 0:one w ;1:neutral ;2:selection ; 3:discrete ;4:freqs ;
                  * 5:gamma ;6:2gamma ;7:beta ;8:beta &w ;9:beta &gamma ;
                * 10:beta &gamma+1 ; 11:beta &normal>1 ; 12:0 &2normal>1 ;
                * 13:3normal>0

icode = 0  * 0:universal code ; 1:mammalian mt ; 2-10:see below
Mgene = 0
             * codon: 0:rates, 1:separate ; 2:diff pi, 3:diff kapa, 4:all diff
             * AA: 0:rates, 1:separate

fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
kappa = 2  * initial or fixed kappa
fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
omega = .4 * initial or fixed omega, for codons or codon-based AAs

fix_alpha = 1  * 0: estimate gamma shape parameter ; 1: fix it at alpha
alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
Malpha = 0  * different alphas for genes
ncatG = 8  * # of categories in dG of NSsites models

getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 1  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

Small_Diff = .5e-6
cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
method = 0  * Optimization method 0: simultaneous ; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Annexe 3 : Fichiers de programmation du logiciel PAML pour la datation moléculaire. Les modèles choisis étant identiques pour les trois gène, les fichiers de commande sont également les mêmes.

Sans horloge moléculaire

```
outfile = mlb      * main result file
noisy = 3   * 0,1,2,3: how much rubbish on the screen
verbose = 0   * 1: detailed output, 0: concise output
runmode = 0   * 0: user tree ;  1: semi-automatic ;  2: automatic
              * 3: StepwiseAddition ; (4,5):PerturbationNNI

model = 4   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
              * 5:T92, 6:TN93, 7:REV, 8:UNREST, 9:REVu ; 10:UNRESTu

Mgene = 0   * 0:rates, 1:separate ; 2:diff pi, 3:diff kapa, 4:all diff

*         ndata = 1
clock = 0   * 0:no clock, 1:clock ; 2:local clock ; 3:CombinedAnalysis
fix_kappa = 0   * 0: estimate kappa ; 1: fix kappa at value below
kappa = 5  * initial or fixed kappa

fix_alpha = 0   * 0: estimate alpha ; 1: fix alpha at value below
alpha = 0.3   * initial or fixed alpha, 0:infinity (constant rate)
     Malpha = 0   * 1: different alpha's for genes, 0: one alpha
ncatG = 5   * # of categories in the dG, AdG, or nparK models of rates
nparK = 0   * rate-class models. 1:rK, 2:rK &fK, 3:rK &MK(1/K), 4:rK &MK

nhomo = 0   * 0  & 1: homogeneous, 2: kappa for branches, 3: N1, 4: N2
getSE = 0   * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0   * (0,1,2): rates (alpha>0) or ancestral states

Small_Diff = 7e-6
cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*       icode = 0  * (with RateAncestor=1. try "GC" in data,model=4,Mgene=4)
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
method = 0  * Optimization method 0: simultaneous ; 1: one branch a time
```

Contraint par l'horloge moléculaire stricte.

```
outfile = mlb       * main result file
noisy = 3   * 0,1,2,3: how much rubbish on the screen
verbose = 0   * 1: detailed output, 0: concise output
runmode = 0   * 0: user tree ;  1: semi-automatic ;  2: automatic
              * 3: StepwiseAddition ; (4,5):PerturbationNNI

model = 4   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
              * 5:T92, 6:TN93, 7:REV, 8:UNREST, 9:REVu ; 10:UNRESTu

Mgene = 0   * 0:rates, 1:separate ; 2:diff pi, 3:diff kapa, 4:all diff

*       ndata = 1
clock = 1   * 0:no clock, 1:clock ; 2:local clock ; 3:CombinedAnalysis
fix_kappa = 0   * 0: estimate kappa ; 1: fix kappa at value below
kappa = 5  * initial or fixed kappa

fix_alpha = 0   * 0: estimate alpha ; 1: fix alpha at value below
alpha = 0.3   * initial or fixed alpha, 0:infinity (constant rate)
Malpha = 0   * 1: different alpha's for genes, 0: one alpha
ncatG = 5   * # of categories in the dG, AdG, or nparK models of rates
nparK = 0   * rate-class models. 1:rK, 2:rK &fK, 3:rK &MK(1/K), 4:rK &MK

nhomo = 0   * 0  & 1: homogeneous, 2: kappa for branches, 3: N1, 4: N2
getSE = 1   * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 1   * (0,1,2): rates (alpha>0) or ancestral states

Small_Diff = 7e-6
cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*       icode = 0  * (with RateAncestor=1. try "GC" in data,model=4,Mgene=4)
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
      method = 0  * Optimization method 0: simultaneous ; 1: one branch a time
```

Annexe 4 : Modèles évolutifs proposés par le logiciel TREEFINDER selon les trois critères d'optimisation AIC, AICc et BIC.

Critère d'optimisation	Gène
	B646L
AIC	TN[{0.4287784,0.075763218,0.075763218,0.075763218,0.075763218,0.26816873},{0.30406935,0.22963042,0.27694084,0.18935939}]:G[{0.26277617}]:5
AICc	HKY[{0.34962292,0.07518854,0.07518854,0.07518854,0.07518854,0.34962292},{0.3148038,0.23579701,0.26863269,0.1807665}]:G[{0.25186975}]:5
BIC	HKY[{0.34962292,0.07518854,0.07518854,0.07518854,0.07518854,0.34962292},{0.3148038,0.23579701,0.26863269,0.1807665}]:G[{0.25186975}]:5
	B646L sans les codons sous pression de sélection positive
AIC	J2[{0.43269429,0.053612579,0.089511168,0.053612579,0.089511168,0.28105822},{0.29615456,0.23041843,0.28632368,0.18710334}]:G[{0.27496356}]:5
AICc	HKY[{0.36433807,0.067830963,0.067830963,0.067830963,0.067830963,0.36433807},{0.30432245,0.23788099,0.27214829,0.18564827}]:G[{0.25854158}]:5
BIC	HKY[{0.36433807,0.067830963,0.067830963,0.067830963,0.067830963,0.36433807},{0.30432245,0.23788099,0.27214829,0.18564827}]:G[{0.25854158}]:5
	B646L : enracinement (séquences acides aminés)
AIC	LG[,{0.061000383,0.011113273,0.056584713,0.044882136,0.048119449,0.063111019,0.025922282,0.070455413,0.048005357,0.07182183, 0.020130492,0.063630577,0.056175098,0.034982076,0.042266792,0.081172837,0.068952034,0.074024696,0.015234422,0.042415119}]:G[{6.0479526}]:5
AICc	betHIV[,{0.06834353,0.0088337151,0.056075296,0.035618185,0.045407932,0.05332492,0.02254033,0.092181707,0.040477983,0.066781116,0.021644275, 0.07441051,0.049570231,0.025703702,0.036312204,0.085995225,0.084851483,0.085470649,0.01274899,0.033708017}]:I[{0.0031730735}]
BIC	betHIV[,{0.06834353,0.0088337151,0.056075296,0.035618185,0.045407932,0.05332492,0.02254033,0.092181707,0.040477983,0.066781116,0.021644275, 0.07441051,0.049570231,0.025703702,0.036312204,0.085995225,0.084851483,0.085470649,0.01274899,0.033708017}]:I[{0.0031730735}]
	E183L sans séquences recombinantes
AIC	HKY[{0.35857465,0.070712673,0.070712673,0.070712673,0.070712673,0.35857465},{0.24761849,0.27203852,0.28738558,0.19295741}]:G[{0.6661414}]:5
AICc	HKY[{0.3,0.1,0.1,0.1,0.1,0.3},{0.24814266,0.26840417,0.28696566,0.19648752}]
BIC	HKY[{0.35853344,0.07073328,0.07073328,0.07073328,0.07073328,0.35853344},{0.24884104,0.27140815,0.28704984,0.19270097}]:G[{0.66254449}]:5
	E186L sans pression sans séquences recombinantes et sans codons sous pression de sélection positive
AIC	HKY[{0.36795375,0.066023127,0.066023127,0.066023127,0.066023127,0.36795375},{0.25208513,0.27217353,0.2839628,0.19177854}]:G[{0.7786304}]:5
AICc	HKY[{0.3,0.1,0.1,0.1,0.1,0.3},{0.25249709,0.26989398,0.28434148,0.19326745}]
BIC	HKY[{0.36795375,0.066023127,0.066023127,0.066023127,0.066023127,0.36795375},{0.25208513,0.27217353,0.2839628,0.19177854}]:G[{0.7786304}]:5

	CP204L	
AIC		J1[{0.43808954,0.058131065,0.058131065,0.091228291,0.091228291,0.26319175},{0.25593531,0.18471455,0.33799931,0.22135083}]:GI[{0.69144422,0.40061967}]:5
AICc		HKY[{0.34905087,0.075474566,0.075474566,0.075474566,0.075474566,0.34905087},{0.2625216,0.20136794,0.32624701,0.20986345}]:G[{0.24385559}]:5
BIC		TN[{0.44404929,0.073583117,0.073583117,0.073583117,0.073583117,0.26161824},{0.24747063,0.1912927,0.33933668,0.2219}]:G[{0.24765114}]:5
	CP204L sans codons sous pression de sélection positive	
AIC		J3[{0.4706562,0.047758471,0.074688934,0.074688934,0.047758471,0.28444899},{0.25571993,0.17470425,0.34752136,0.22205446}]:G[{0.26162436}]:5
AICc		HKY[{0.37614478,0.061927611,0.061927611,0.061927611,0.061927611,0.37614478},{0.26931495,0.18452492,0.33577374,0.21038639}]:G[{0.25654386}]:5
BIC		HKY[{0.37614478,0.061927611,0.061927611,0.061927611,0.061927611,0.37614478},{0.26931495,0.18452492,0.33577374,0.21038639}]:G[{0.25654386}]:5

Annexe 5 : Article traitant de la classification et des origines du virus PPA. Cet article a été soumis à la revue Plos One

Title: Comprehensive phylogenetic reconstructions of African swine fever virus: proposal for a new classification and molecular dating of the virus

Author list: Michaud, V. and Albina, E.

Abstract: African swine fever (ASF) is a highly lethal disease of domestic pigs caused by the only known DNA arbovirus. It was first described in Kenya in 1921 and since then a substantial number of isolates have been collected worldwide. However, only few phylogenetic studies have been carried out to better understand the relationships between the isolates. In this paper, comprehensive phylogenetic reconstructions were made using publicly and newly generated sequences of hundreds ASFV isolates of the last 70 years. Analyses focused on B646L, CP204L and E183L genes from 356, 251 and 123 isolates, respectively. Phylogenetic analyses were achieved using maximum likelihood and Bayesian coalescence methods and a new lineage based nomenclature is proposed to designate 35 different clusters. In addition, dating of ASFV origin was carried out from the molecular data sets. To avoid bias, diversity due to positive selection or recombination events was neutralized. The molecular clock analyses revealed that ASFV strains currently circulating have evolved over 300 years, with a time to the most recent common ancestor (TMRCA) going back to the early 18th century.

Keywords: African swine fever – phylogeny – molecular clocking – B646L – E183L – CP204L

INTRODUCTION

African swine fever (ASF) is an infectious and contagious hemorrhagic disease of domestic pigs (Penrith & Vosloo 2009). It is highly lethal, causing up to 100% mortality in naive animals with devastating effects on pig production and animal trade, and major economic losses in affected countries (Costard *et al.* 2009). First described by Montgomery in 1921 in Kenya (Montgomery 1921), ASF spread rapidly from eastern Africa to most of sub-Saharan countries where it has often become endemic. From Africa, it reached Europe, i.e. Portugal in 1957 and again in 1960 from where it colonized Spain, France and Belgium. From there, the virus reached Latin America during the 70s-80s. In Europe, It remained endemic in the Iberian Peninsula up to the middle of the 90s and the disease is still present in Sardinia (Costard *et al.* 2009). Recently, the disease has been re-introduced on the borders of Europe, in Georgia in 2007 (Rowlands *et al.* 2008) and then extended to the Caucasus and Russia (Gulenkin *et al.* 2011). No vaccine is available and disease control is only based on quarantine and animal slaughtering. In this context, its large ability to spread makes ASF virus one of the most important infectious threats for the domestic pig industry worldwide. African swine fever virus (ASFV) is a large icosahedral and enveloped dsDNA virus, the unique recognized DNA arbovirus and also the unique member of the *Asfarviridae* family and *Asfivirus* genus (Dixon 2005). However, ASFV shares characteristics with the other members of the *Nucleo-cytoplasmic Large DNA virus* family (Ogata *et al.* 2009) suggesting that they all may have had a common ancestor (Iyer 2006).

ASFV is supposed to be an ancestral virus of soft tick (*Ornithodoros* genus) (Plowright 1977) infecting wild swine like warthogs (*Phaecochoerus aethiopicus*), bushpigs (*Potamochoerus porcus*) and giant forest hog (*Hylochoerus meinertzhageni*) with asymptomatic effects. The virus replicates in ticks and then is transmitted to wild swine during blood feeding, the wild life being considered as natural reservoir of the virus. The virus can persist in ticks for years, even in quiescent ticks waiting for host feeding. The sylvatic cycle of ASFV established between ticks and wild suids can be maintained indefinitely. This cycle allows the maintenance of virus circulation and probably enables the persistence of ancient viruses and the emergence of new variants as well. At the laboratory level, virus variants were initially characterized by genome size and enzymatic restriction profiles. A high level of variability is observed mainly within the 35 kb at the 3' end and 15 kb at the 5' end of the genome (170-190 kb). These two regions contain the multigene families (MGF) which vary in number between isolates and enable virus variability by gene homologous recombination. Moreover, variability is also generated by a change in the number of amino-acid repeats in 14 proteins including the envelop protein p54 encoded by the E183L gene (Sun *et al.* 1995). More recently, gene sequencing and analysis were introduced to increase differentiation between ASFV isolates collected worldwide. Two groups (Bastos *et al.* 2003, Lubisi *et al.*, 2005) used phylogenetic reconstructions to discriminate ASFV isolates. Based on partial sequence of B646L gene coding for the major viral protein (MCP) VP72, their trees showed a very close relation between West African, European and south American isolates, all clustered in genotype 1. Despite more than 50 years of circulation among three continents, a limited accumulation of genetic changes has made impossible the discrimination of isolates within the genotype 1. In contrast, eastern and southern African isolates are more diverse and segregate into 21 additional genotypes. This could be explained by the fact that these viruses are propagated within a sylvatic cycle in contrast to viruses of genotypes 1 that mainly replicated in domestic pigs, although they were secondarily detected in European soft ticks *O. erraticus* and wild boars in Spain and Portugal. This accredits the assumption that the virus diversity may be generated during the sylvatic cycle of the virus (Dixon & Wilkinson 1988). Other genes or genome sequences have been used successfully to discriminate ASFV isolates collected at a regional level: for instance the B602L gene from the central variable region of the genome (CVR, coding J9L protein), the CP204L (coding the phospho-protein P32) and E183L (envelop protein p54) genes have been used to further split the local isolates (Nix *et al.* 2006) (Gallardo *et al.* 2009 ; Rowlands *et al.* 2008).

The aim of this study was to reassess phylogenetic reconstructions and nomenclature of ASFV including more recent sequences and to explore the evolution of the virus based on a comprehensive analysis of the available sequence datasets. Accordingly, three genes were targeted, all of them being the most sequenced and uploaded in public databases. B646L, E183L and CP204L genes belong to the most conserved central part of the genome and encode the structural virus proteins VP72 (capsid), p54 (membrane protein) and p32 (membrane protein), respectively. They are also known to generate antibodies in pig (Neilan *et al.* 2004). The origin and the evolution of the virus were inferred from these three genes.

MATERIALS AND METHODS

Data set

A large collection of ASFV isolates were included in this study. The majority of ASFV sequences used in this study were downloaded from GenBank (http://www.ncbi.nlm.nih.gov) and a CISA-INIA web site data bank (http://wwwx.inia.es/cisa/asfv/). Additional sequences of Madagascar isolates were generated after virus isolation on pig alveolar macrophages from pig sampled during outbreaks between 1998 and 2008. These sequences are interesting for the study of ASFV evolution since they are considered to have derived from a unique introduction of the virus in 1998: twenty-one samples were selected to cover the whole territory and the period 1998-2010. PCRs were performed using the following primers: VP72-d (5'-GGCACAAGTTCGGACATGT-3') and VP72-U (5'-GTACTGTAACGAAGCAGCACAG-3'), E183L for p54 (5'-GGTTGGTTTTCAAATGTTGGCGAAGGTA-3') and E183Lrev p54 (5'-CCATAAATTCTGTAATTTCATTGCGCCACAAC-3') and p30/32-P1 (5'-TG CCAAGCATACATAAGTTG-3') and p30/32-P2 (5'-ATTT TGCTGTTTATGAATCC-3') for the amplification of B646L, E183L and CP204L genes, respectively. PCR products were cloned in E coli and sequences generated from these clones by a private company (Cogenics, France). Sites with mutations were particularly checked for sequencing errors: only bases confirmed by two direction reading were retained as mutations. In all, the analyses were performed on 356 sequences (399 nt long), 251 sequences (480 nt long) and 123 sequences (543 nt long) for B646L, E183L and CP204L genes, respectively.

Phylogenetic inference

Sequence analysis

Sequences were aligned by ClustalW with default parameters and then scrutinized and edited using Mega version 5 software (Tamura K. 2011). From the multiple sequence alignments, an index of substitution saturation to estimate the degree of sequence information was calculated using Dambe software (Xia & Xie 2001). DNA polymorphism was also analyzed. The site diversity between two sequences (π) and the number of segregating sites (i.e. the number of sites where one or several substitutions occurred) were obtained by DnaSP version 5 software (Librado & Rozas 2009). In the segregating sites, the ratio of transitions and transversions was assessed. The average number of nucleotide differences (k) between two sequences was also determined. The maximum and minimum rates of substitutions observed from the data set were established manually using an Excel spreadsheet. All this information was used initially to check the quality of the sequences and secondarily as calibration priors for molecular clock analyses.

Test for recombination

The presence of sequence recombination events in the data set was assessed in the multiple alignments with RDP3 package version 3 (Heath *et al.* 2006) using the default setting for all recombination tests applied on linear sequences (RDP (Martin & Rybicki 2000), GENECONV (Padidam *et al.* 1999), MAXCHI (Smith 1992), BOOTSCAN/RESCAN (Martin *et al.* 2005), and SISCAN (Gibbs *et al.* 2000)).

Phylogenetic reconstruction

Maximum likelihood reconstructions (Felsenstein 1981), generating trees that best fit the evolution of a set of sequences through a probabilistic model of evolution, were performed using TREEFINDER version March 2011 software (Jobb *et al.* 2004). The selection of the evolution model was done according to the Akaike information criterion (AIC)(Akaike 1974), the corrected AIC (AICc) (Sugiura 1978) and Bayesian Information Criterion (BIC)(Schwarz 1978) with a number of gamma rate categories fixed at 5. The consensus model given by the three information criteria or alternatively, the simplest model was selected for the reconstruction. Thus, the B646L tree was constructed under HKY + Γ_5 model (Hasegawa *et al.* 1985 ; Yang 1994). For E183L and CP204L, HKY + Γ_5 and HKY on one hand and HKY + Γ_5 and TN+ Γ_5 models on the other hand were selected. The most complex model GTR (Rodriguez *et al.* 1990) was also systematically included and compared with the others. All the reconstructions were done on 1000 replicates and bootstraps were approximated using the Expected-Likelihood Weights defined by Strimmer and Rambaut (2002) applied on local rearrangements (LR-ELW) as implemented in TREEFINDER.

Bayesian inference phylogeny was performed using Monte Carlo Markov Chain (MCMC) implemented in MrBayes version 3.1 software (Huelsenbeck & Ronquist 2001 ; Ronquist & Huelsenbeck 2003). According to the best fit models proposed by TREEFINDER, MrBayes was set with HKY + Γ_5, HKY and HKY + Γ_5, and HKY+ Γ_5, for B646L, E183L and CP204L, respectively. GTR model was also performed for each gene. MCMC was run for a maximum of 10 millions of trees or alternatively when the run reached stationarity as measured by a standard deviation of split frequencies ever becoming lower than 0.01 or fluctuating randomly above 0.01 for at least 500,000 generated trees. Consensus trees were generated after having discarded the first 25% of the MCMC burn-in phase.

Tree congruence with data sets was tested by submitting them to the statistical test ELW (Strimmer & Rambaut 2002) implemented in TREEFINDER. The tree selected for each gene was the one with the highest ELW score.

Since ASFV is the only member of the *Asfarviridae* family, outgroup viruses for tree rooting were selected in the closest related DNA virus families *Poxviridae*, *Iridoviridae* and *Phycodnavirodae* (Iyer *et al.* 2001). Four viruses were included: Lymphocystis disease virus 1 (LDV1) and Invertebrate iridescent virus 6 (CIV6) for Iridoviridae, Vaccinia virus for Poxviridae and Ectocarpus siliculosus virus (EsV) for Phycodnaviridae. Because a high level of nucleotide divergence, multiple sequence alignments were done on the complete amino-acid sequences of the major capsid protein of both outgroup viruses and ASFV isolates (equivalent to B646L protein) using Mega5 software.

Tree reconstructions were performed using maximum likelihood method set with LG + I model on 1000 replicates. The topology of the resulting rooted tree was subsequently applied for placing roots on the B646L, E183L and CP204L trees.

Analysis of selection pressure

Codons under positive selection pressure in DNA coding sequences may evolve faster than the natural evolutionary rate of the virus genome. In order to avoid bias in the molecular clocking analysis, the selection pressure acting on the targeted genes was assessed. The ratio of non synonymous (dN) – synonymous substitution (dS) per site (dN/dS ratio) was calculated and the codons under positive selection pressure identified by using Codeml software implemented in PAML 4 package.

ASFV genotyping

Isolate genotyping was assessed by comparing the genetic distance between all B646L sequences. Average intra- and inter-branch distances were globally compared in order to determine the strength of cluster segregation. Additionally, a haplotype network of the isolates was constructed using TCS1.21 software to identify relationships between isolates potentially poorly represented by conventional phylogenetic tree reconstruction. At last, specific nucleotide signatures of the different ASFV clusters were searched using multiple sequence alignments containing only the 67 unique B646L sequences. The three approaches were finally combined to raise conclusions on ASFV genotyping.

Molecular dating

Two methods were performed in parallel and compared in order to determine the evolutionary rate and the time to the most recent common ancestor (Tmrca) of circulating ASFV isolates. The first was based on the maximum likelihood method Baseml implemented in PAML 4 package (Yang 2007) and the second on the Bayesian MCMC implemented in BEAST package version 1.6.2 (Drummond & Rambaut 2007). Codons under positive selection (dN/dS > 1) and recombined sequences were removed from the multiple sequence alignments to avoid bias in the substitution rate determination and consequently in the Tmrca estimation. The best fit tree generated in the phylogenetic reconstructions was used to perform Baseml implemented in PAML 4 package, using as evolution models, the HKY + Γ_5 for B646L, CP204L and E183L genes. Strict and relaxed molecular clock hypotheses (Zuckerkandl & Pauling 1965) were used to generate dated trees for all genes. These two trees were individually compared with the tree generated without a clock constraint to accept or reject the molecular clock hypothesis. A likelihood ratio test (LRT) and a χ^2 comparison was performed to support this analysis. For the relaxed molecular clock, branches delineating the different genotypes were individually relaxed.

All analyses performed with BEAST package were done under an uncorrelated lognormal relaxed clock model. Considering that at least 20% of our sequences were from isolates persisting in the wild life, a constant population size prior was selected. The initial value and the range of substitution rates estimated in the previous sequence analyses were entered into the model of evolution. For each gene, analyses of two independent runs of 100 million steps were performed with 1/10000 trees sampled. MCMC samples were examined using Tracer version 1.4 (Rambaut A. 2003), the first 25% samples of the chain being discarded as burn-in phase. Tree consensus was generated using the maximum clade credibility (MCC) tree using Tree Annotator version 1.4.7 (Drummond & Rambaut 2007). Only posterior probabilities higher than 0.90 were indicated.

Tree visualization

All trees were represented edited in Fig Tree version 1.3.1 developed by Andrew Rambaut (http://tree.bio.ed.ac.uk/software/figtree/).

RESULTS

Comprehensive phylogenetic inference of ASFV depicts 4 major lineages

Before phylogenetic inference, data sets and multiple sequence alignments were thoroughly examined to eliminate misalignments and ensure correct framing of coding sequences. All gaps were considered as missing information to avoid artificial nucleotide divergence. None of the different methods used in RDP3 package, could identify recombination events in B646L and CP204L sequences. In contrast, several recombination events were detected among E183L sequences. A total of 17 isolates were subsequently removed from the E183L multiple sequence alignments: 16 were Italian isolates and one was a South African isolate. The recombination events were all identical for Italian isolates (Figure 1). There were no saturated codons in our alignments (DAMBE, p_{value}<<0.03), thus supporting the fact that the genetic information in the data sets was convenient for phylogenetic studies.

To check the nucleotide composition of the alignments, statistical tests were performed using DnaSP software. The tests gave the number of nucleotide substitutions, the average diversity per site between two sequences (π) and the average nucleotide difference between two sequences (k). Two genes (B646L and CP204L) showed approximately a half-lower diversity compared to the third one (Table 1). In addition, the latter show a clear bias in non-synonymous mutations. Based on the observed nucleotide substitutions, the minimum and the maximum evolutionary rates were also calculated from each multiple alignments (Table 1). *dN/dS* of each gene was determined (Table 1) as well as the amino acids under positive selection in the alignments. This led to remove 6 nt (2 aa) from B646L alignment, 9 nt (3 aa) from CP204L alignment and 27 nt (9 aa) from E183L alignments for subsequent molecular dating analyses.

Table 1: Summary of the multiple sequence alignments analyses

Gene	Size, in nt	Polymorphic	Syn.	Non	π	k	dN/dS	μ observed

		sites		Syn.				
B646L	399	110 (27.6%)	62	60	0.03416	13.629	1.253	[0.0053 - 0.14]
B646L w/oP	393							[0.07 - 0.002]
E183L	480	263 (54.7%)	88	214	0.07283	34.960	1.158	[0.0115 - 0.215]
E183L w/o P	453	215 (51.2%)						[0.0075-0.166]
CP204L	543	154 (28.2%)	83	88	0.06738	36.588	8.42	[0.00557 - 0.102]
CP204L w/o P	534	146 (27.3%)						[0.097 – 0.0054]

π: average diversity per site between two sequences (number of nucleotide differences per site between two sequences). k: average number of nucleotide difference between two sequences. μ: substitution rate. w/o P: without codons under positive selection.

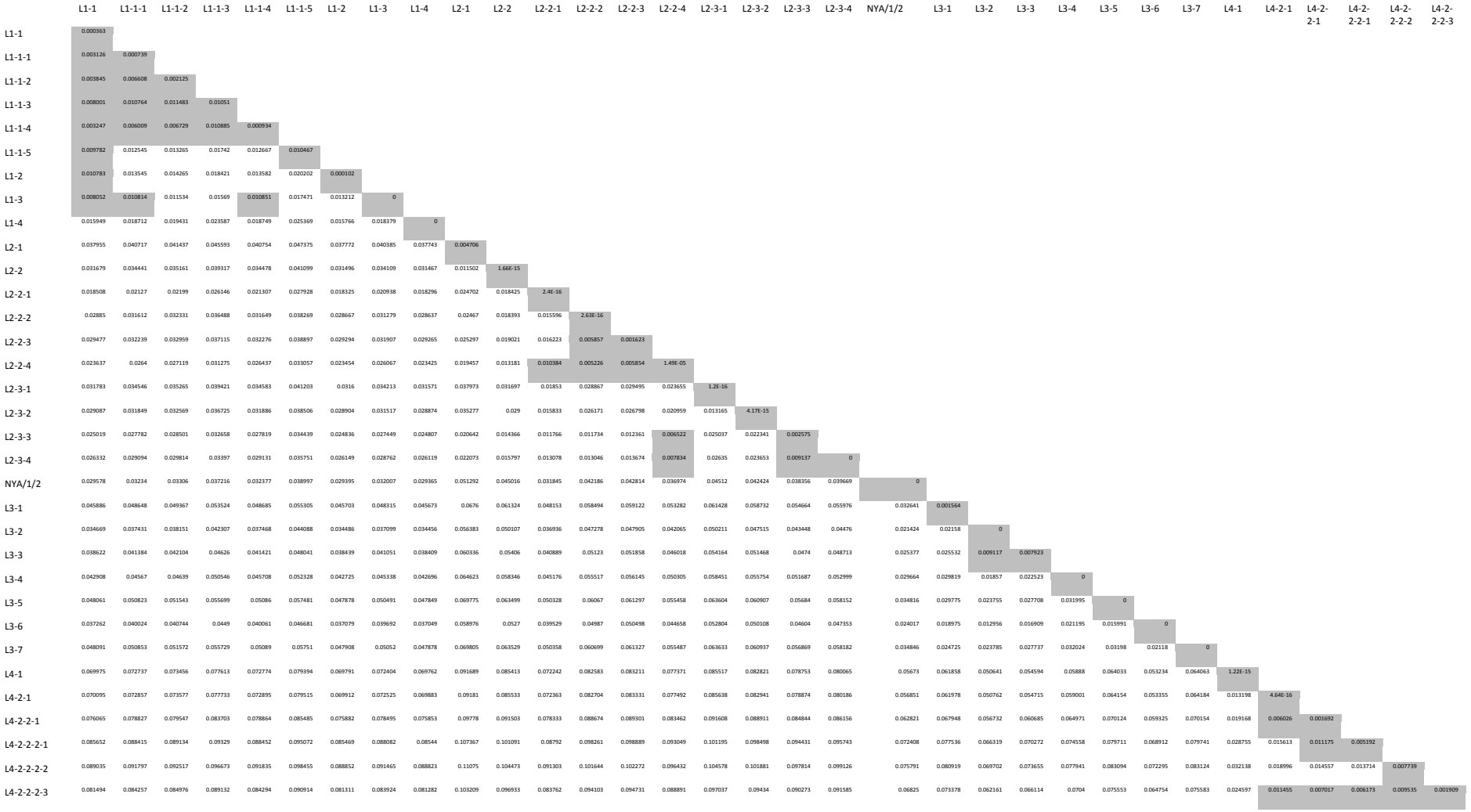
The rooted tree constructed from the multiple sequence alignments of the major capsid protein amino acid sequences of ASFV and four out-grouped viruses from three other virus families showed that the common ancestor of all these viruses connects the ASFV group within Eastern African isolates, more precisely between the genotype VIII on one hand and genotypes IX and X on the other hand (Figure 2). Accordingly, the root on all subsequent trees was placed in this position. This reconstruction also shows that the *Asfarviridae* family is rather divergent from the three other families.

Phylogenetic trees constructed with B646L sequences show four major lineages (L) (Figure 3): L1 includes the previously described genotypes I, II, XVII and XVIII, and L2, the genotypes III, IV, V, VI, VII, XIX, XX, XXI, XXII and an ungenotyped isolate (Cro3.5). L3 includes genotypes VIII, XI, XII, XIII, XV, XVI and one isolate TAN/08/MAZIMBU, previously included within the genotype XV (Misinzo *et al.* 2010). L4 gathers genotypes IX and X. Interestingly, the NYA/1/2 isolate ascribed to the genotype XIV, is the only isolate that does not segregate within one of the four lineages. However, bootstrap value of its branch is <70%, thus rendering difficult final conclusion on this isolate. Further clustering of the isolates within these four lineages becomes tricky because of the presence of long branches and multifurcation for some isolate groups or sub-lineages. The TCS network analysis showed that conventional phylogenetic reconstruction based on bifurcations may fail to explain the complex relationships between some isolates (Figure 4). The TCS network confirms the existence of the four lineages that include the same isolates as in bifurcated reconstructions. However, it seems to better explain the relationships of isolates within a given genotype (e.g. genotype I or X) or between distinct genotypes (e.g. between genotypes III, IV, XIX, XX, XXI or between genotypes IX and X). In these cases the pattern of isolates relationships is not strictly bifurcative. Three ways exist between genotype XIX and genotype XX: through genotypes III or IV and/or XXI, which represent internal nodes of the tree, and two ways between genotypes IX and X. Within genotype X, several isolates are internal nodes of the tree, meaning that an isolate can have more than one ancestor, which is inconsistent with bifurcative relationships between isolates. In attempts to refine the clusterization of ASFV isolates, the multiple sequence alignments containing 67 unique B646L sequences were searched for specific molecular signatures (Figure 5). Lineage 1 is characterized by 2 nt and L2, L3 and L4 by 4, 6 and 12 nt, respectively. The genotype XIV, which is not included in one of the four lineages, is characterized by 8 nt (G88, G93, G162, T214, C240, T258, T333, and T348). However, this is the only virus generating this branch which in addition is not supported by a high bootstrap value (<70%). Therefore, it cannot yet be considered as a fifth lineage. Lineages can be subsequently sub-divided into sub-lineages: 4 for the lineage 1, 3 for the lineage 2, 7 for the lineage 3 and 2 for the lineage 4 (Figure 5). Further sub-divisions can be drawn from the molecular signatures (Figure 5) and all are supported by the evolutionary distance matrix except for some sub-lineages within L1-1, L1-2, L1-3, L2-2, L2-3 and L4-2-2 (Table 2). The average evolutionary distance inside and between all sub-lineages were 0.0023 and 0.055, respectively. L4 is the most complex lineage, composed of isolates from countries of the Great Lakes Region in Africa (Tanzania, Uganda, Burundi and Kenya) and divided into several sub-lineages. The sub-lineage L4.2 (including isolates belonging to the former genotype X) is the most diverse with isolates clustering into seven sub-lineages (from L4-2-1 to L4-2-2-2-3). This new clusterization into lineages almost perfectly overlays the previous genotype discrimination with the exception of Cro3.5 isolate which forms a new cluster within L2 (sub-lineage L2-3-4) and TAN/O8/MAZIMBU isolate which splits from genotype XV to form a new sub-lineage of L3 (L3-7).

The trees generated with CP204L and E183L genes (data not shown) confirmed the existence of four lineages including the same genotypes. However, the E183L gene tree shows some differences in the clustering: SPEC/205 belonged to L1.1.1 lineage with B646L while it moves to L3.2.2.3 (genotype XI) with E183L. NYA/1/2, the sole member of the former genotype XIV in B646L classification and which segregated between lineages L1, L2 and L3 is replaced within the lineage L3 in E183L classification. Whether these modifications may be ascribed to inter-gene recombination events remains unclear.

Molecular dating leads to a most recent common ancestor of about 300 years old

The strict molecular clock hypothesis, meaning an equal substitution rate for every nucleotide sites all along the DNA sequences, was rejected for all three genes using the maximum likelihood analysis performed by Baseml in PAML software suite. In PAML, the branches were individually relaxed in the tree submitted to the analysis. Several trees with different numbers of relaxed branches were tested. The resulting TMRCAs for B646L, E183L, CP204L genes were highly variable: from 1597 BC to 700 AD or even undetermined date (because of a tree likelihood value equal to zero at the beginning of the analysis). This high level of heterogeneity in the TMRCA using maximum likelihood method leads us to select Bayesian approach in the BEAST package. The Bayesian MCMC inference of the three data sets performed with BEAST package showed a satisfactory convergence in the posterior statistic estimates of the substitution rate. Initial value of the parameter substitution/site/year was set at the minimum of μ observed (see table 1). Accordingly, the prior distribution of this parameter was set at 0.1xminimum μ observed and 1xmaximum μ observed. Thus, calibration of molecular clocks were set at $5.3x10^{-3}$ substitution/site/year [$5.3x10^{-4}$ – $1.4x10^{-1}$] for B646L gene, $1.03x10^{-2}$ [$1.03.10^{-4}$ – $3.5.10^{-1}$] for E183L gene and $5.36x10^{-3}$ [$5.36x10^{-4}$ – $1.99x10^{-1}$] for CP204L gene. With these priors, the mean estimates of substitution rate for each gene were finally calculated by BEAST and ranged



	L1-1	L1-1-1	L1-1-2	L1-1-3	L1-1-4	L1-1-5	L1-2	L1-3	L1-4	L2-1	L2-2	L2-2-1	L2-2-2	L2-2-3	L2-2-4	L2-3-1	L2-3-2	L2-3-3	L2-3-4	NYA/1/2	L3-1	L3-2	L3-3	L3-4	L3-5	L3-6	L3-7	L4-1	L4-2-1	L4-2-2-1	L4-2-2-1	L4-2-2-2	L4-2-2-3
L1-1	0.000363																																
L1-1-1	0.003126	0.000739																															
L1-1-2	0.003845	0.006608	0.002125																														
L1-1-3	0.008001	0.010764	0.011483	0.01051																													
L1-1-4	0.003247	0.006009	0.006729	0.010885	0.000934																												
L1-1-5	0.009782	0.012545	0.013265	0.01742	0.012667	0.010467																											
L1-2	0.010783	0.013545	0.014265	0.018421	0.013582	0.020202	0.000102																										
L1-3	0.008052	0.010814	0.011534	0.01569	0.010851	0.017471	0.013212	0																									
L1-4	0.015949	0.018712	0.019431	0.023587	0.018749	0.025369	0.015766	0.018379	0																								
L2-1	0.037955	0.040717	0.041437	0.045593	0.040754	0.047375	0.037772	0.040385	0.037743	0.004706																							
L2-2	0.031679	0.034441	0.035161	0.039317	0.034478	0.041099	0.031496	0.034109	0.031467	0.011502	1.66E-15																						
L2-2-1	0.018508	0.02127	0.02199	0.026146	0.021307	0.027928	0.018325	0.020938	0.018296	0.024702	0.018425	2.4E-16																					
L2-2-2	0.02885	0.031612	0.032331	0.036488	0.031649	0.038269	0.028667	0.031279	0.028637	0.02467	0.018393	0.015596	2.63E-16																				
L2-2-3	0.029477	0.032239	0.032959	0.037115	0.032276	0.038897	0.029294	0.031907	0.029265	0.025297	0.019021	0.016223	0.005857	0.001623																			
L2-2-4	0.023637	0.0264	0.027119	0.031275	0.026437	0.033057	0.023454	0.026067	0.023425	0.019457	0.013181	0.010384	0.005226	0.005854	1.49E-05																		
L2-3-1	0.031783	0.034546	0.035265	0.039421	0.034583	0.041203	0.0316	0.034213	0.031571	0.037973	0.031697	0.01853	0.028867	0.029495	0.023655	1.2E-16																	
L2-3-2	0.029087	0.031849	0.032569	0.036725	0.031886	0.038506	0.028904	0.031517	0.028874	0.035277	0.029	0.015833	0.026171	0.026798	0.020959	0.013165	4.17E-15																
L2-3-3	0.025019	0.027782	0.028501	0.032658	0.027819	0.034439	0.024836	0.027449	0.024807	0.020642	0.014366	0.011766	0.011734	0.012361	0.006522	0.025037	0.022341	0.002575															
L2-3-4	0.026332	0.029094	0.029814	0.03397	0.029131	0.035751	0.026149	0.028762	0.026119	0.022073	0.015797	0.013078	0.013046	0.013674	0.007834	0.02635	0.023653	0.009137	0														
NYA/1/2	0.029578	0.03234	0.03306	0.037216	0.032377	0.038997	0.029395	0.032007	0.029365	0.051292	0.045016	0.031845	0.042186	0.042814	0.036974	0.04512	0.042424	0.038356	0.039669	0													
L3-1	0.045886	0.048648	0.049367	0.053524	0.048685	0.055305	0.045703	0.048315	0.045673	0.0676	0.061324	0.048153	0.058494	0.059122	0.053282	0.061428	0.058732	0.054664	0.055976	0.032641	0.001564												
L3-2	0.034669	0.037431	0.038151	0.042307	0.037468	0.044088	0.034486	0.037099	0.034456	0.056383	0.050107	0.036936	0.047278	0.047905	0.042065	0.050211	0.047515	0.043448	0.04476	0.021424	0.02158	0											
L3-3	0.038622	0.041384	0.042104	0.04626	0.041421	0.048041	0.038439	0.041051	0.038409	0.060336	0.05406	0.040889	0.05123	0.051858	0.046018	0.054164	0.051468	0.0474	0.048713	0.025377	0.025532	0.009117	0.007923										
L3-4	0.042908	0.04567	0.04639	0.050546	0.045708	0.052328	0.042725	0.045338	0.042696	0.064623	0.058346	0.045176	0.055517	0.056145	0.050305	0.058451	0.055754	0.051687	0.052999	0.029664	0.029819	0.01857	0.022523	0									
L3-5	0.048061	0.050823	0.051543	0.055699	0.05086	0.057481	0.047878	0.050491	0.047849	0.069775	0.063499	0.050328	0.06067	0.061297	0.055458	0.063604	0.060907	0.05684	0.058152	0.034816	0.029775	0.023755	0.027708	0.031995	0								
L3-6	0.037262	0.040024	0.040744	0.0449	0.040061	0.046681	0.037079	0.039692	0.037049	0.058976	0.0527	0.039529	0.04987	0.050498	0.044658	0.052804	0.050108	0.04604	0.047353	0.024017	0.018975	0.012956	0.016909	0.021195	0.015991	0							
L3-7	0.048091	0.050853	0.051572	0.055729	0.05089	0.05751	0.047908	0.05052	0.047878	0.069805	0.063529	0.050358	0.060699	0.061327	0.055487	0.063633	0.060937	0.056869	0.058182	0.034846	0.024725	0.023785	0.027737	0.032024	0.03198	0.02118	0						
L4-1	0.069975	0.072737	0.073456	0.077613	0.072774	0.079394	0.069791	0.072404	0.069762	0.091689	0.085413	0.072242	0.082583	0.083211	0.077371	0.085517	0.082821	0.078753	0.080065	0.05673	0.061858	0.050641	0.054594	0.05888	0.064033	0.053234	0.064063	1.22E-15					
L4-2-1	0.070095	0.072857	0.073577	0.077733	0.072895	0.079515	0.069912	0.072525	0.069883	0.09181	0.085533	0.072363	0.082704	0.083331	0.077492	0.085638	0.082941	0.078874	0.080186	0.056851	0.061978	0.050762	0.054715	0.059001	0.064154	0.053355	0.064184	0.013198	4.64E-16				
L4-2-2-1	0.076065	0.078827	0.079547	0.083703	0.078864	0.085485	0.075882	0.078495	0.075853	0.09778	0.091503	0.078333	0.088674	0.089301	0.083462	0.091608	0.088911	0.084844	0.086156	0.062821	0.067948	0.056732	0.060685	0.064971	0.070124	0.059325	0.070154	0.019168	0.006026	0.001692			
L4-2-2-2-1	0.085652	0.088415	0.089134	0.09329	0.088452	0.095072	0.085469	0.088082	0.08544	0.107367	0.101091	0.08792	0.098261	0.098889	0.093049	0.101195	0.098498	0.094431	0.095743	0.072408	0.077536	0.066319	0.070272	0.074558	0.079711	0.068912	0.079741	0.028755	0.015613	0.011175	0.005192		
L4-2-2-2-2	0.089035	0.091797	0.092517	0.096673	0.091835	0.098455	0.088852	0.091465	0.088823	0.11075	0.104473	0.091303	0.101644	0.102272	0.096432	0.104578	0.101881	0.097814	0.099126	0.075791	0.080919	0.069702	0.073655	0.077941	0.083094	0.072295	0.083124	0.032138	0.018996	0.014557	0.013714	0.007739	
L4-2-2-2-3	0.081494	0.084257	0.084976	0.089132	0.084294	0.090914	0.081311	0.083924	0.081282	0.103209	0.096933	0.083762	0.094103	0.094731	0.088891	0.097037	0.09434	0.090273	0.091585	0.06825	0.073378	0.062161	0.066114	0.0704	0.075553	0.064754	0.075583	0.024597	0.011455	0.007017	0.006173	0.009535	0.001909

Table 2: Estimates of evolutionary distances between ASF lineages and sub-lineages. The average intra-sublineage diversity was 0.0023, whereas the average inter-sub-lineage was 0.055. In the matrix, diversity lesser than 5x(intra-sub-lineage diversity)= 0.0115 are shown in grey boxes: all sub-lineages (Lx-x) differ from the others by a higher diversity, except for some sub-lineages within L1.1, L1.2, L1.3, L2.2, L2.3 and L4.2.2.

from $2.7x10^{-4}$ (E183L) to $6.49x10^{-4}$ (B646L) subst/site/year (Table 3). These results are robust in terms of clock model, rate distribution and population size parameters. The dated trees generated four lineages as previously described (Figure 6) and again, the same isolates were found within these lineages. The four lineages were organized in the same way for B646L and E183L genes: L1 and L2 on one hand and L3 and L4 on the other hand. In contrast, CP204L gene tree rendered different connections: L1, L2 and L4 together and L3 on the other hand. In all cases, the oldest lineage (Tmrca = 111 years) was L4 that gathers isolates from Eastern Africa, the presumed birthplace of ASFV. It was followed by L1 (104 years), L2 (74 years) and L3 (47 years).

Table 3: Summary results of all tests done in BEAST for molecular clocking, models, evolution rates and the TMRCA obtained.

gene	clock model	population size	evolutionary model	LRT	mean rate subst/site/year [95% HPD]	TMRCA [95% HPD]
B646L	strict	constant	HKY + Γ_5	-1993.8	1.131×10^{-4} [7,6 x 10-5 – 1,5 x 10^{-4}]	1622 [1466 – 1768]
	UCLN	constant	HKY + Γ_5	-1593.4	6.9×10^{-4} [5.3×10^{-4} – 9.13×10^{-4}]	1712 [1894 – 1465]
CP204L	strict	constant	HKY + Γ_5	-2453.07	6.76×10^{-5} [$1,9 \times 10^{-5}$ – $1,2 \times 10^{-4}$]	850 [-204 – 1610]
	UCLN	constant	HKY + Γ_5	-2398.03	6.6×10^{-4} [5.57×10^{-4} – 8.75×10^{-4}]	1700 [1422 – 1891]
E183L	strict	constant	HKY + Γ_5	-3902.7	1.76×10^{-4} [$9,1 \times 10^{-5}$ – $2,6 \times 10^{-4}$]	1529 [1243 – 1745]
	UCLN	constant	HKY + Γ_5	-2161.9	2.7×10^{-4} [2.07×10^{-4} – 4.03×10^{-4}]	1426 [1422 – 1720]

HPD = highest posterior density

Discussion

Because of the localization of the major capsid protein VP72 in the virus core preventing exposition to circulating neutralizing antibodies, the corresponding B646L gene is not expected to be submitted to the immune system pressure. Accordingly, only two amino-acid positions were detected as being under positive selection, suggesting no real impact on the evolutionary force. Therefore, the rate of substitutions of the VP72 is probably bearing relevant information to estimate the natural virus evolution. The VP72 homologues of closely related virus families has been previously used in evolutionary studies (Tidona *et al.* 1998) and for a decade in ASFV phylogenetic reconstructions. In contrast, P54 is an envelope protein and the pressure of the immune system on E183L evolution is revealed by nine amino-acid positions placed under positive selection and/or recombination events within the gene sequence. P32 is also an envelope protein but is involved in translation of viral genes by its interactions with hnRNP cellular protein (Hernaez *et al.* 2008). In this context, mutations may be detrimental and thus, the gene submitted to purifying selection as corroborated by the detection of only three amino-acid positions placed under positive selection.

The evolution of ASFV mapped through partial genomic sequences and phylogenetic reconstructions show a certain degree of complexity that may not be well represented by bifurcative methods. However, both bifurcative and network analyses in this study clearly gave a clear clusterization into four major lineages (L1 to L4) while only three have been described so far (Boshoff *et al.* 2007). Within these lineages, molecular signatures of the twenty-two already described genotypes were established and two new sub-lineages can be proposed, represented by Cro3.5 isolate and TAN/08/MAZIMBU previously ascribed to genotype XV. Molecular signatures do not rely on the same number of substitutions and do not have an equal weight. For instance, L1 is characterized by 2 specific nucleic acid positions and 4 synonymous substitutions while L4 is defined by 12 sites and 13 nucleotides substitutions of which 3 are not synonymous. Within the L1 lineage, the genotype I which is the most represented in terms of sequences (Europe, West Africa, Caribbean and South America) is characterized by only one synonymous substitution (A216). This mutation leads however to an increase in ASFV codon preference for Alanine (GCG to GCA) (http://www.kazusa.or.jp/codon/) which has surely helped to fix the substitution in the lineage for almost 60 years in three continents. Besides the molecular signature, the distance matrix also supports our proposal for new ASFV classification which includes the previous genotype subdivision and additional sub-clustering.

ASFV shows a high evolutionary rate compared to the substitution rate of other DNA viruses (Duffy *et al.* 2008). Consequently, this high mutation rate led to very recent TMRCAs: the most common ancestor of ASFV strains currently circulating was determined to have emerged in around three centuries, in 1700. It is commonly agreed that ASFV is native from east Africa as the disease was first described in Kenya 1921 after a first outbreak in 1907. Then, ASFV showed a great ability to spread worldwide during decades following major trade routes. In the wild, the virus is thought to be originally a virus of tick (Plowright 1977) as it infects argasids ticks of the *Ornithodoros* genus. *Ornithodoros* which infest warthogs' burrows are endophile ticks meaning that they need regular temperature and hygrometry. They also are photophobic so they do not really spread out over long distances. ASFV is transmitted horizontally and vertically between ticks (Hess *et al.* 1989 ; Plowright *et al.* 1974) and between ticks and juvenile wild swine that stay in and close to their burrows. Under such circumstances, the virus is not supposed to spread much and its genetic drift over long periods may have resulted in isolated spots of diversity maintained by the sylvatic cycle with only few entries of new strains. In contrast, the domestic pig cycle is short with a dead-end disease essentially transmitted by contacts with pig or pig meat and rarely by tick bites. Accordingly, phylogenetic trees constructed in this study showed increased diversity within lineages of eastern and southern African isolates submitted to a sylvatic cycle compared to others from domestic pigs of other regions. New variants are not easy to characterize because of the lack of sequence data from their parent lineage. For example, TAN/08 Mazimbu isolate collected in Tanzania in 2008 and originally placed in genotype XV (Misinzo *et al.* 2010) constitutes in this study a sub-lineage of L3. Thus, it should not be considered as a re-emergence of the TAN/01/1 isolate collected during an outbreak in Tanzania in 2001. Sixteen Italian isolates showed recombination event in the E183L gene and were subsequently removed from the corresponding reconstruction. This does not change the affiliation of these isolates to L1 as demonstrated by B646L and CP204L reconstructions (not shown). However, since all these isolates are linked together and show the same recombination events, assuming they all have emerged from a common recombined ancestor, the possibility that they will form a new sub-lineage within L1 has to be considered.

Three different genes and two methods have been used to consolidate TMRCA estimation. Maximum likelihood method using PAML package showed that a strict molecular clock could not be validated for our set of genes. However, it did not provide consistent results when using a relaxed clock, with TMRCAs from -12000 to 1500. In contrast, the Bayesian approach generated consistent results. B646L and CP204L analyses both dated a TMRCA around 1700 AD while E183L set it up three centuries earlier (1400 AD) with a rate of subst/site/year estimated to be about 2 times smaller for the latter. This lower rate of subst/site/year for E183L estimated by BEAST combined with an observed higher number of synonymous and non-synonymous substitutions may explain a higher TMRCA for this gene. As illustrated by the dN/dS analysis, the role of the immune system on sequence variability may have influenced the sequence evolution of this gene and consequently rendered a backward TMRCA. Therefore, the natural evolution of the virus may be better represented by B646L and CP204L genes. The TMRCA scale going back to 1700 AD for all ASFV isolates can be considered with confidence since within this scale, the TMRCA of lineage L1-1 and L1-2 were 1943/1955 (for B646L/CP204L genes) and 1990 (for both genes), respectively. L1-1 is supposed to have emerged in the late 1950s (Bastos *et al.* 2003) and L1-2 includes mainly isolates from Madagascar that were first introduced in 1998 (Gonzague *et al.* 2001). The substitution rates determined in this study are much higher than expected comparing to other large dsDNA viruses like gamma-herpes viruses of vertebrate (10^{-9} subs/site/year) or even small dsDNA viruses like the John Cunningham polyomavirus (10^{-7} subs/site/year) (Duffy *et al.* 2008). With a substitution rate comprised between 10^{-4} and 10^{-5}, ASFV is approaching RNA viruses which usually have 10^{-2} to 10^{-5} subs/site/year (Hanada *et al.* 2004).

Like many other large dsDNA viruses (Holmes 2004), ASFV may have co-evolved with its host. This means a long and ancient history of the virus in the wild. High substitution rate combined with recent TMRCAs is not consistent with ancient co-evolution of viruses and their hosts which in contrast should lead to a low rate of substitution (Holmes & Drummond 2007). However, for a virus that replicates at high level in its host, a low rate of subst/site/replication can still lead to an increased accumulation of diversity which in turns generates high rates of subst/site/year (Hughes *et al.* 2009). This has been described for highly contagious viruses that induce acute forms of infection and show a higher observed rate of subst/site/year (Firth *et al.* 2010). In contrast, an asymptomatic infection of the host may not allow an exponential replication rate. ASFV presents these two characteristics, being asymptomatic in natural African wild swine and soft ticks while highly contagious and lethal in domestic pigs. Consequently, a stochastic event may have occurred around 300 years from now to explain the emergence of a common ancestor to all known ASFV isolated so far in domestic and wild pigs. Our assumption is based on the introduction of domestic pigs in Africa. Domestic pigs have a Eurasian and North African ancestral wild boar origins (Gifford-Gonzalez 2011). Despite one archaeological record of pig introduction in South Africa between the 3rd and 7th centuries (Plug 2001) domestic pigs were not present in Eastern and Southern African livestock because of the nomadic lifestyle of pastoralists at this time (Swart 2010). Domestic pigs may have been brought first by Chinese around 600 years ago (Levathes 1994) then by Portuguese 300 to 400 years ago (Blench 1999), both during their exploration and conquest period for trade opportunities. The assumption of pig introduction from Europe and Far East was confirmed by phylogenetic analysis revealing contributions of both origins in the genetic pattern of local African pigs (Ramirez *et al.* 2009). Following the circumnavigation of Africa by European nations during 15th - 17th centuries, pig breed types were introduced during 16th and 17th centuries (Swart 2010), mainly brought by Portuguese to East African coast via Goa. Pig breeding diffused then slowly northward from Mozambique (Blench 1999). Portuguese did not colonized Kenya for settlement but as a step to India and definitely left the country in 1720 after being defeated by Arabs in 1698. Despite Arab colonization and pig eating taboo, domestic pigs were still eaten by ethnic groups like Waata present in southern Kenya since the 16th century and called Walyankuru: "those who eat pig" (Kusimba 2000). This may have enabled virus to spread silently among sensitive pig species. Kenya was then colonized by British. Extensive pig industry in the native region of ASFV started after a massive loss of bovine cattle due to rinderpest outbreak at the end of the 19th century. Pigs were massively imported for breeding by colonizers from Seychelles in 1904 and from England in 1905. Pig farming was free ranging at this time and the first outbreak of ASF was reported in 1907. Trade routes and virus resistance in the environment then enabled further spreading of ASFV.

Acknowledgement:

The authors wish to acknowledge la Direction de la Santé Animale et du Phytosanitaire du Ministère de l'Agriculture, de l'Elevage et de la Pêche of Madagascar for the permission to use Madagascar isolates in this study. For facilitating this work in Madagascar, Tantely Randriamparany, François Roger and Renaud Lancelot are also warmly thanked. This work was financially supported by Wellcome Trust (N°210183. 183, AHDW/03/04), the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement

KBBE- 211691- ASFRISK, and also partially funded by the European Union through the Network of Excellence EPIZONE (number FOOD-CT-2006-016236)

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

Bastos AD, Penrith ML, Cruciere C*, et al.* (2003) Genotyping field strains of African swine fever virus by partial p72 gene characterisation. *Arch Virol* **148**, 693-706.

Blench RM (1999) A history of pigs in Africa. In: Blench, R.M., Mac Donald, K., editors. Origins and development of African livestock: archaeology, genetics, linguistics and ethnography. *Florence, K.Y.: Routledge Books.*, 335-367.

Boshoff CI, Bastos AD, Gerber LJ, Vosloo W (2007) Genetic characterisation of African swine fever viruses from outbreaks in southern Africa (1973-1999). *Vet Microbiol* **121**, 45-55.

Costard S, Wieland B, de Glanville W*, et al.* (2009) African swine fever: how can global spread be prevented? *Philos Trans R Soc Lond B Biol Sci* **364**, 2683-2696.

Dixon LK, Escriban, J.M., Martins, C., Rock, D.L., Salas, M.L., Wilkinson, P.J. (2005) Asfarviridae. In: Fauquet, CM.M, Mayo, M.A., Maniloff, J., Deselberger, U., Ball, L.A., editors. *Virus Taxonomy, VIIIth report of the ICTV. London (UK): Elsevier/Academic Press*, 135-143.

Dixon LK, Wilkinson PJ (1988) Genetic diversity of African swine fever virus isolates from soft ticks (Ornithodoros moubata) inhabiting warthog burrows in Zambia. *J Gen Virol* **69 (Pt 12)**, 2981-2993.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214.

Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267-276.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-376.

Firth C, Kitchen A, Shapiro B*, et al.* (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol* **27**, 2038-2051.

Gallardo C, Mwaengo DM, Macharia JM*, et al.* (2009) Enhanced discrimination of African swine fever virus isolates through nucleotide sequencing of the p54, p72, and pB602L (CVR) genes. *Virus Genes* **38**, 85-95.

Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.

Gifford-Gonzalez DaH, O. (2011) Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *J World Prehist* **24**, 1-23.

Gonzague M, Roger F, Bastos A*, et al.* (2001) Isolation of a non-haemadsorbing, non-cytopathic strain of African swine fever virus in Madagascar. *Epidemiol Infect* **126**, 453-459.

Gulenkin VM, Korennoy FI, Karaulov AK, Dudnikov SA (2011) Cartographical analysis of African swine fever outbreaks in the territory of the Russian Federation and computer modeling of the basic reproduction ratio. *Prev Vet Med* **102**, 167-174.

Hanada K, Suzuki Y, Gojobori T (2004) A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol* **21**, 1074-1080.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174.

Heath L, van der Walt E, Varsani A, Martin DP (2006) Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol* **80**, 11827-11832.

Hernaez B, Escribano JM, Alonso C (2008) African swine fever virus protein p30 interaction with heterogeneous nuclear ribonucleoprotein K (hnRNP-K) during infection. *FEBS Lett* **582**, 3275-3280.

Hess WR, Endris RG, Lousa A, Caiado JM (1989) Clearance of African swine fever virus from infected tick (Acari) colonies. *J Med Entomol* **26**, 314-317.

Holmes EC (2004) The phylogeography of human viruses. *Mol Ecol* **13**, 745-756.

Holmes EC, Drummond AJ (2007) The evolutionary genetics of viral emergence. *Curr Top Microbiol Immunol* **315**, 51-66.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755.

Hughes AL, Irausquin S, Friedman R (2009) The evolutionary biology of poxviruses. *Infect Genet Evol* **10**, 50-59.

Iyer LA, Balaji, S., Koonin, E.V. and Aravind, L. (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research* **117**, 156-184.

Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* **75**, 11720-11734.

Jobb G, von Haeseler A, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**, 18.

Kusimba CMaK, S.B. (2000) Hinterlands and cities: Archaeological investigations of economy and trade in Tsavo, sout-eastern Kenya. *Department of Anthropology, The field Museum, 1400 S. Lake Shore Drive, Chigogo, Illinois, USA, 606005* **54**, 13-24.

Levathes LE (1994) When China Ruled the Seas: The Treasure Fleet of the Dragon Throne, 1405-1433. *New York: Oxford University Press.*

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.

Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98-102.

Misinzo G, Magambo J, Masambu J*, et al.* (2010) Genetic characterization of African swine fever viruses from a 2008 outbreak in Tanzania. *Transbound Emerg Dis* **58**, 86-92.

Montgomery R (1921) On a form of swine fever occurring in British East Africa (Kenya colony). *J. Comp. Pathol.* **34**, 159, 191, 243-262.

Neilan JG, Zsak L, Lu Z*, et al.* (2004) Neutralizing antibodies to African swine fever virus proteins p30, p54, and p72 are not sufficient for antibody-mediated protection. *Virology* **319**, 337-342.

Nix RJ, Gallardo C, Hutchings G, Blanco E, Dixon LK (2006) Molecular epidemiology of African swine fever virus studied by analysis of four variable genome regions. *Arch Virol* **151**, 2475-2494.

Ogata H, Toyoda K, Tomaru Y*, et al.* (2009) Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virol J* **6**, 178.

Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.

Penrith ML, Vosloo W (2009) Review of African swine fever: transmission, spread and control. *J S Afr Vet Assoc* **80**, 58-62.

Plowright W (1977) Vector transmission of African swine fever virus. In: *Seminar on Hog cholera, classical swine fever and African swine fever*, pp. 575-587. Eur 5904EN, commission of the European communities.
Plowright W, Perry CT, Greig A (1974) Sexual transmission of African swine fever virus in the tick, Ornithodoros moubata porcinus, Walton. *Res Vet Sci* **17**, 106-113.
Plug IaB, S. (2001) The distribution of macrommals in Southern Africa over the past 30,000 years. *Transvaal Museum Monograph.* **113, South Africa**.
Rambaut A. D, A.J. (2003) Tracer [computer program]. *http://evolve.zoo.ox.ac.uk/software/*.
Ramirez O, Ojeda A, Tomas A*, et al.* (2009) Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol Biol Evol* **26**, 2061-2072.
Rodriguez F, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. *J Theor Biol* **142**, 485-501.
Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.
Rowlands RJ, Michaud V, Heath L*, et al.* (2008) African swine fever virus isolate, Georgia, 2007. *Emerg Infect Dis* **14**, 1870-1874.
Schwarz G (1978) *Estimating the dimension of a model*. *Ann. Stat.* **6**, 461-464.
Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129.
Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci* **269**, 137-142.
Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communication in Statist. Theor. Meth.* **7**, 13-26.
Sun H, Jacobs SC, Smith GL, Dixon LK, Parkhouse RM (1995) African swine fever virus gene j13L encodes a 25-27 kDa virion protein with variable numbers of amino acid repeats. *J Gen Virol* **76 (Pt 5)**, 1117-1127.
Swart T, Kotze, A., Olivier, P.A.S. and Grobler, J.P. (2010) Microsatellite-based characterization of Southern African domestic pigs (*Sus scrofa domestica*). *South African Journal of Animal Science* **40**, 121-132.
Tamura K. PD, Peterson N., Stecher G., Nei M., and Kumar S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **10**, 2731-2739.
Tidona CA, Schnitzler P, Kehm R, Darai G (1998) Is the major capsid protein of iridoviruses a suitable target for the study of viral evolution? *Virus Genes* **16**, 59-66.
Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**, 371-373.
Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**, 306-314.
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591.
Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357-366.

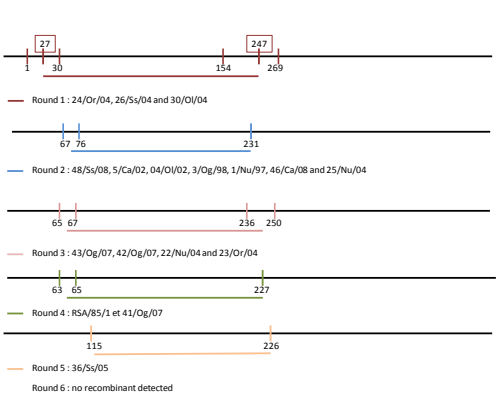


Figure 3: localization of recombination events detected in E183L sequence alignment. 16 Italian isolates and 1 South African isolate were detected to be recombinant. Italian isolates are linked and the fact that recombination events take place in the same region of the sequence lead to think that these isolates emerged from a common ancestor.

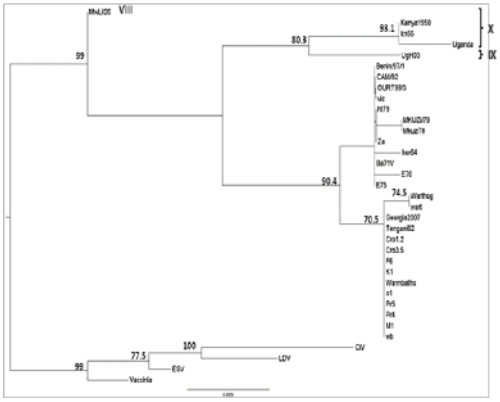


Figure 2: Rooted tree constructed from amino-acid multiple alignments of the major capsid protein of ASFV isolates and four out-grouped viruses. The tree was constructed under a LG + I model and maximum likelihood method with 1,000 bootstrap resampling. Numbers indicate the statistical value (Expected-Likelihood Weight) of internal nodes, given in percentages (only numbers above 70% are indicated). The out-groups connect ASFV group by the branch conducting from genotype VIII (MwLil20/1) to genotypes IX (UgH03) and X (Kenya1950, kn66 and Uganda).

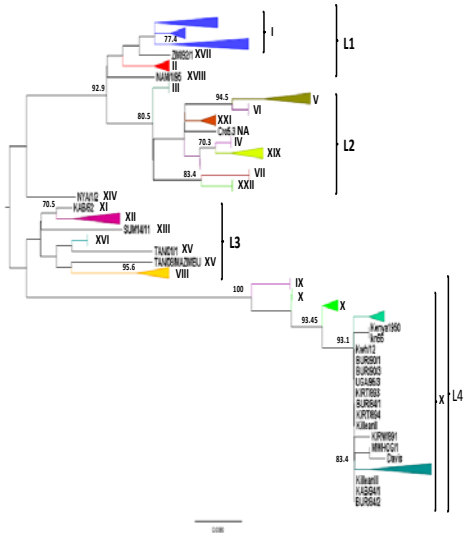


Figure 3: B646L gene phylogenetic tree describing ASFV relationships. The tree was constructed under HKY85 + Γ5 evolutionary model with 1,000 bootstrap resampling. Numbers indicate the statistical value (Expected-Likelihood Weight) of internal nodes, given in percentages (only numbers higher than 70 are indicated). Lineages were collapsed for improved tree visibility. The tree shows four main lineages (L1 to L4).

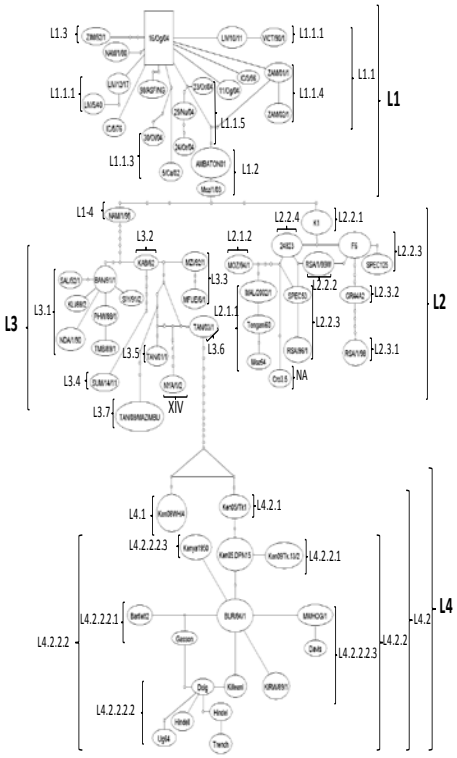
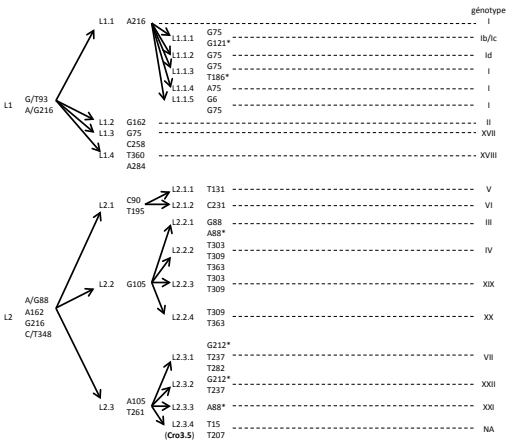


Figure 4: Haplotype network constructed with TCS software. The network shows the same four main lineages that were observed in the bifurcative phylogenetic tree constructed in maximum likelihood under the HKY+ Γ5 model, but clearly demonstrates that relationships between some ASFV isolates are too complex to be resolved by only bifurcations.



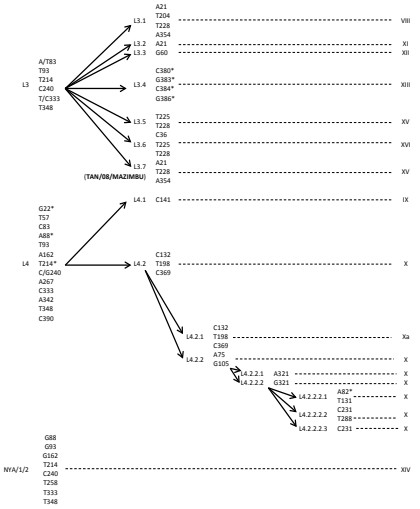


Figure 5: Molecular signatures of ASFV lineages and sub-lineages. Corresponding genotypes are indicated in the right column. Non synonymous substitutions are labeled with "*". NA: non assigned.

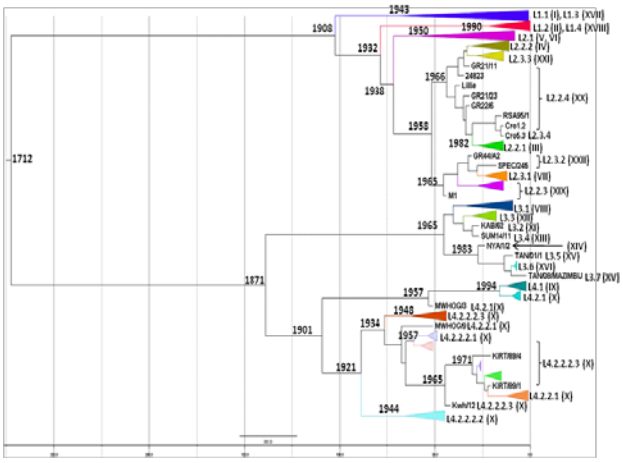


Figure 6 : Dated tree representing evolutionary relationships of B646L gene between ASFV isolates. Lineages and corresponding genotypes are indicated. The tree was constructed by Beast software and the MCMC were run for 10^8 times. Time for all isolatesTMRCA is 1712. TMRCAs of lineage L1-1 (genotype I) and lineage L1-2 (genotype II) are 1943 and 1990, respectively.